Weakly Supervised Learning by a Confusion Matrix of Contexts

William Wu

Faculty of Engineering and Information Technology,
University of Technology Sydney, Australia
William.Z.Wu@student.uts.edu.au

Abstract. Context consideration can help provide more background and related information for weakly supervised learning. The inclusion of less documented historical and environmental context in researching diabetes amongst Pima Indians uncovered reasons which were more likely to explain why some Pima Indians had much higher rates of diabetes than Caucasians, primarily due to historical, environmental and social causes rather than their specific genetic patterns or ethnicity as suggested by many medical studies.

If historical and environmental factors are considered as external contexts when not included as part of a dataset for research, some forms of internal contexts may also exist inside the dataset without being declared. This paper discusses a context construction model that transforms a confusion matrix into a matrix of categorical, incremental and correlational context to emulate a kind of internal context to search for more informative patterns in order to improve weakly supervised learning from limited labeled samples for unlabeled data.

When the negative and positive labeled samples and misclassification errors are compared to "happy families" and "unhappy families", the contexts constructed by this model in the classification experiments reflected the Anna Karenina principle well - "Happy families are all alike; every unhappy family is unhappy in its own way", an encouraging sign to further explore contexts associated with harmonizing patterns and divisive causes for knowledge discovery in a world of uncertainty.

Keywords: Weakly supervised learning, context, confusion matrix, context construction, contextual analysis.

1 Introduction

In many real-world data mining tasks, such as healthcare data investigation and financial profile classification, it is often the case that not all sample records in training sets are fully labeled due to various constraints such as cost and capability restriction, sampling and classification limitation. The weakly supervised learning approach is meant to handle these situations adaptively when only a small number of training samples are labeled by initial probing classifications, and then some forms of self-learning are engaged to revise progressively and to label the rest accordingly.

Weakly supervised learning may have emanated from the "bootstrapping" way of semi-supervised self-learning in the field of data mining and classification [1,2] and many strategies and techniques have been developed, for example, active learning by uncertainty sampling, query-by-committee and a decision-theoretic approach assisted by human experts as "oracles" [3], semi-supervised learning by an expectation–maximization approach, co-training and multi-view learning via separation of attribute sets with integration of result learning [4], and reinforcement learning by an iteration of an agent's action and reward functions in a Markov decision process [5,6]. Collectively speaking, the more information available on the samples, the more informed decisions can be made by the experts as "oracles" and by the researchers to improve the self/co-training rulesets and the action-reward learning logic.

As an example of improved learning by considering contexts around the available data, the inclusion of less documented historical and environmental context in researching diabetes amongst Pima Indians by Schulz el al. and Phihl [7,8], uncovered reasons which were more likely to explain why some Pima Indians "have up to ten times the rate of diabetes as Caucasians" which were not due to their specific genetic patterns or ethnicity, but because of their loss of access to water from the Gila River in the 1900s which adversely affected their agricultural supply and farm work, and their later adaption of modern Western fast food and life-style. In comparison, another group of Pima Indians from the same tribe migrated to Mexico in the 1890s and continued their traditional farm work and traditional diet, have a significant lower rate of diabetes despite these two groups sharing a common genetic background [8,9].

A confusion matrix is a simple and effective way to summarize classification results and algorithm performance in a tabular form by tallying the category results, making result comparison easy between class label categories [10-12]. When using the term "incremental context" to describe how data samples are sorted in ascending order to emulate the value growth path of a lead attribute, and the term "correlational context" to describe the subsequent and correlated value variation of other related attributes along the value growth path of the lead attribute, then the term "categorical context" can describe the cross-category interrelation when each matrix cell hosts both "incremental context" and "correlational context" for their respective result category, which makes cross-category context comparison simple, visual and systematic, and facilitates learning in a category-by-category and side-by-side manner.

2 Context Construction with Confusion Matrix

2.1 Post-Classification Analysis and Data Normalization for Easy Comparison

Context construction can start as a part of the post-classification analysis process based on the initial set of training samples which have been labeled, and the underlying classification platform can be based on any classification algorithm, e.g. decision tree, neural network or naive Bayes. The resulting confusion matrix can be evaluated first and if suitable, be transformed into a table with each of its category cells to house the incremental and correlational contexts of their respective training samples.

In order to construct incremental and correlational context for an easy and consistent comparison between attributes in a dataset, the values of different attributes measured in different units are converted into one standardized value range between 0 and 100 via a min-max normalization routine, and ten key value growth stages of each attribute can then be established by ten deciles based on these standardized values.

2.2 Construction of Categorical, Incremental and Correlational Context

Key steps in this construction process include:

1. Evaluate initial classification result and its confusion matrix
2. Define a selection list of lead attributes and their rotation/significance order
3. Select a lead attribute and redistribute its data records with standardized values into categorized subsets according to their confusion matrix result categories
4. For each categorized subset sorted by the current lead attribute
   4a. Extract the median record from each decile to simulate ten value growth stages in a vertical way within its categorized cell as incremental context
   4b. Start from the current lead attribute and for its ten growth stages, connect and plot lines across other attributes accordingly as correlational context
   4c. Evaluate these incremental and correlational contexts in the form of line and value patterns between matrix categories as categorical context
5. Select the next lead attribute from the rotation list determined in Step 2 and repeat Steps 3 & 4 to construct more contextual models for further analysis

This context construction process starts as a multi-contextual model with categorical, incremental and correlational context. It is hoped that this is only the beginning, and other factors can later be introduced to this multi-contextual model to expand the scope of data exploration and knowledge discovery for weakly supervised learning.

3 Experiments and Analysis

3.1 Context Construction for Attribute “length” in the Page Blocks Dataset

Experiments on the labeled training sets are based on ten UCI datasets [13] but only two are highlighted in this short paper. WEKA’s [14] C4.5/J48 decision tree is utilized as the initial probing classifier, but other algorithms can also be used.

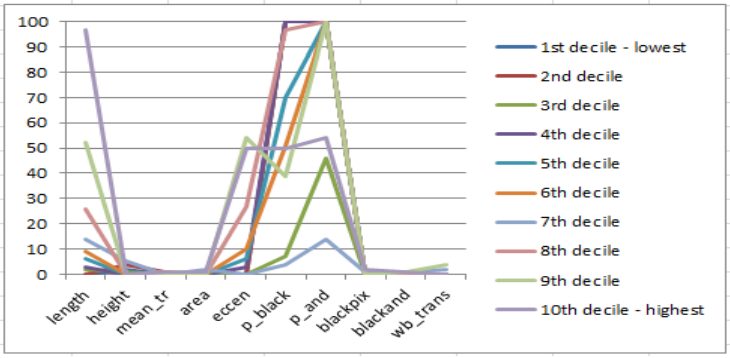
When testing on the Page Blocks dataset, the attribute “length” is selected as the first lead attribute because it is ranked as the most significant attribute by the gain ratio evaluator, and its multi-contextual model is shown in Table 1. The four cells in Table 1 represent the four confusion matrix result categories based on an initial binary classification for the labeled text or non-text samples. The y-axis in each cell represents the ten key value growth stages of attribute “length”, and the x-axis represents the other related attributes in the training set. The ten plots starting from “length” and connecting across the other attributes represent how the values of other attributes

from the same key growth stage samples change according to the incremental growth of the "length" value in a correlational way, showing how the converging and diverging value patterns differentiate between the four confusion matrix categories and reflecting the Anna Karenina principle [15] - text type page blocks share the same regular characteristics, non-text type page blocks and classification errors differ in various ways due to numerous factors such as varying color pigment and text font-size, etc. Context models constructed for other attributes also present similar patterns.

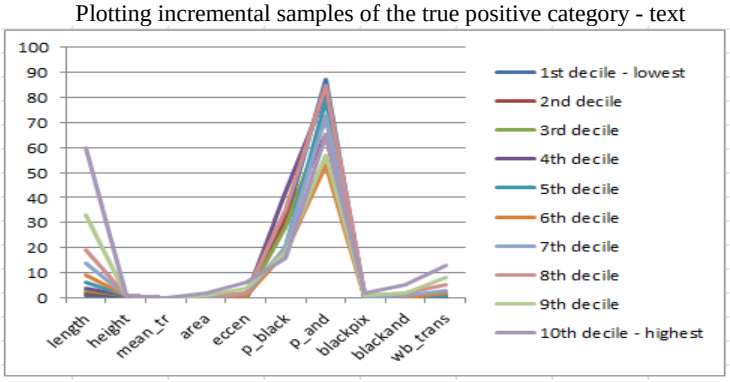
Table 1. A confusion matrix of contexts constructed for the Page Blocks dataset

Visualization and comparison of a confusion matrix of incremental, correlational and categorical context using "length" as the lead attribute

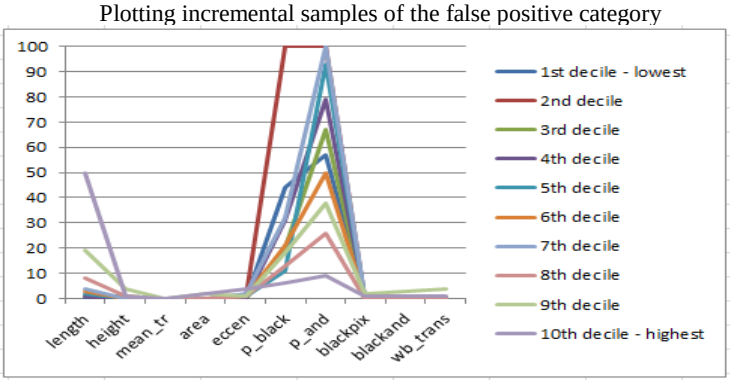
Plotting incremental samples of the true negative category – non-text



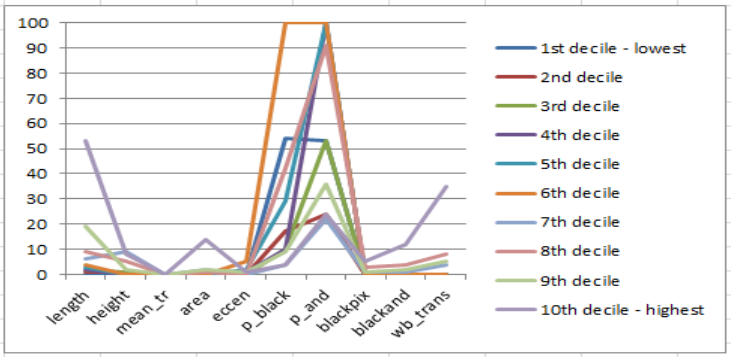
Plotting incremental samples of the true positive category - text



Plotting incremental samples of the false positive category



Plotting incremental samples of the false negative category



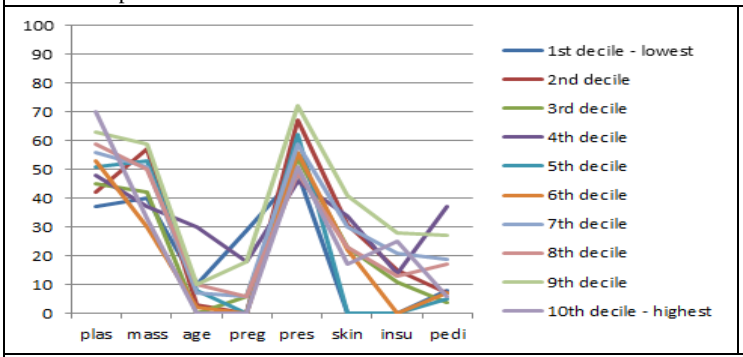
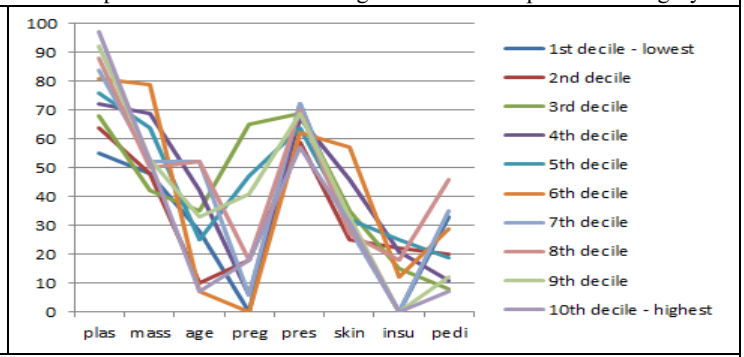
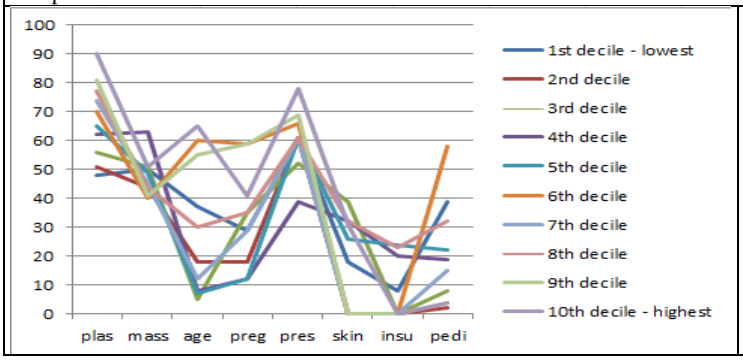
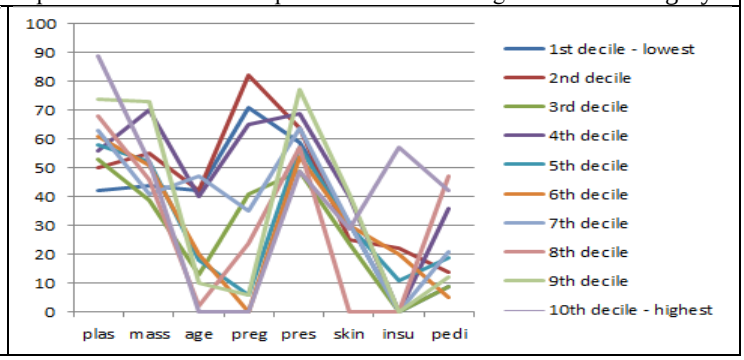
3.2 Context Construction for Attribute "plas" in the Pima Diabetes Dataset

In this test case demonstration of the Pima Indians diabetes dataset, the attribute "plas" is selected as the lead attribute and its multi-contextual model is shown in Table 2. One distinctive pattern is that the majority of "plas" values in the true negative category are bounded in the middle range between the normalized median value of 37 and 70, and as the "plas" value increases, the values of the other attributes vary rather uniformly, and most are bounded by a narrow value range. In contrast, "plas" values in the true positive category start from a higher point of 55 and scatter in a wider val-

ue range up to 97, and the values of the other related attributes, e.g. “mass”, “age” and “pedigree”, also fluctuate in a wider value range with higher starting points. The two error categories, false positives and false negatives, share similar “bigger and rougher” value variation patterns similar to the true positive category.

In essence, the true negative samples show more uniformity than the true positive samples and the error samples. This is a theme that matches the contrast between the benefits of harmonized patterns and the burdens of unrestrained value fluctuations signified by the Anna Karenina principle and shared by other attributes and datasets.

Table 2 - Matrix of incremental, correlational & categorical context for the Pima dataset

Comparison of the incremental and correlational context for the lead attribute “plas” between the “true negative” and “true positive” category	
	
Comparison of the incremental and correlational context for the lead attribute “plas” between the “false positive” and “false negative” error category	
	

4 Potential Issues and Conclusion

Context construction by coupling incremental sampling with cross-attribute plotting under a confusion matrix structure may sound trivial, and its representation of incremental, correlational and categorical context may look inconsequential. However, the focus of this study is not about breaking and winning in complexity, it is about sorting and connecting data elements to gain more understanding about their internal relationship within context and to initiate such contextual analysis with simplicity and practicality, hence the inclusion of ROC curve and F-score will be considered next.

A more specific issue concerns the suitability and adequacy of using ten sorted deciles to represent ten key value growth stages of a lead attribute. While systematic sampling can be considered as a kind of stratification with equal sampling fractions, its value growth representation can be over-optimistic and will need improvement.

In conclusion, the experiments and examination of this multi-contextual model can indeed highlight certain converging and diverging patterns in a contextual and meaningful way, and they may potentially be applicable to unlabeled data for pattern-matching and error-estimation when learning from labeled samples. It may seem far-fetched, but supported by some interesting findings, it is worth exploring.

References

1. Nadeau, D. and Sekine, S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), pp.3-26 (2007).
2. Zhou, Z.H. A Brief Introduction to Weakly Supervised Learning. National Science Review, 5(1), pp.44-53 (2017).
3. Settles, B. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), pp.1-114 (2012).
4. Zhu, X. Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison, 2(3) (2006).
5. Sutton, R.S. and Barto, A.G. Introduction to Reinforcement Learning. Cambridge: MIT press (1998).
6. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. and Petersen, S. Human-level control through deep reinforcement learning. Nature, 518(7540), p.529 (2015)
7. Schulz, L.O., Bennett, P.H., Ravussin, E., Kidd, J.R., Kidd, K.K., Esparza, J. and Valencia, M.E. Effects of Traditional and Western Environments on Prevalence of Type 2 Diabetes in Pima Indians in Mexico and the US. Diabetes Care, 29(8), pp.1866-1871 (2006).
8. Phihl, T. Medical Geography of the Pima Indian Reservation Diabetes Epidemic: The Role of the Gila River. Undergraduate Honors Theses. Paper 304 (2012).
9. Valencia, M.E., Weil, E.J., Nelson, R.G., Esparza, J., Schulz, L.O., Ravussin, E. and Bennett, P.H. Impact of lifestyle on prevalence of kidney disease in Pima Indians in Mexico and the United States. Kidney International, 68, pp. S141-S144. (2005)
10. Fawcett, T. An Introduction to ROC Analysis. Pattern recognition letters, 27(8), pp.861-874. (2006).
11. Visa, S., Ramsay, B., Ralescu, A.L. and Van der Knaap, E. Confusion Matrix-based Feature Selection. In MAICS, pp. 120-127 (2011).
12. Patel, K., Bancroft, N., Drucker, S.M., Fogarty, J., Ko, A.J. and Landay, J. Gestalt: Integrated Support for Implementation and Analysis in Machine Learning. In Proceedings of the 23nd annual ACM symposium, pp. 37-46. (2010).
13. Bache, K., Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science, Irvine, CA (2013).
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1 (2009).
15. Diamond, J.M. Guns, germs and steel: a short history of everybody for the last 13,000 years. Random House (1998).