Similarity Measure of Breast Cancer Datasets Using Fuzzy Rule-Based Classification by Attribute

Tengyue Li, University of Macau, Macau
Simon Fong, University of Macau, Macau

ABSTRACT
To compare with two datasets based on attributes by using classification algorithms, for the attributes, the authors need to select them by rules and the system is known as rule-based reasoning system which classifies a given test instance into a particular outcome from the learned rules. The test instance carries multiple attributes, which are usually the values of diagnostic tests. In this article, the authors propose a classifier ensemble-based method for comparison of two breast cancer datasets. The ensemble data mining learning methods are applied to rule generation, and a multi-criterion evaluation approach is used for selecting reliable rules over the results of the ensemble methods. The efficacy of the proposed methodology is illustrated via an example of two breast cancer datasets. This article introduces a novel fuzzy rule-based classification method called FURIA, to obtain a relationship between two breast cancer datasets. Hence, it can find the similarity between these two datasets. The new method is compared vis-à-vis with other classical statistical approaches such as correlation and mutual information gain.

KEYWORDS
Attributes, Breast Cancer, Classification Algorithms, Comparison, FURIA, Similarity

INTRODUCTION
Data analysis is becoming more and more important and widely adopted nowadays. Many areas use data analysis in biomedical industry that plays a significant role in our society. It is known that biomedical analysis involves a huge amount of data, and often being characterized with many attributes. The impact of successful biomedical analysis is meaningful to mankind and helpful in improving healthcare applications in our society. In our research, we attempted grouping similar biomedical data, but the difference is that we group these attributes by rules, and these rules can guide us to distinguish the useful attributes among the rest. In a combined dataset, we can compare the useful attributes and find the common attributes to obtain a quantitative similarity measure of two different datasets.

The learning of rule-based classification models has been an active area of research for a long time. In fact, the interest in rule induction goes far beyond the field of machine learning itself and includes other fields, notably fuzzy systems (Hüllermeier, 2009). This is hardly surprising, given that rule-based models have always been a cornerstone of fuzzy systems and a central aspect of research in that field. To a large extent, the popularity of rule-based models can be attributed to their comprehensibility, a distinguishing feature and key advantage in comparison to many other (black-
box) classification models. Despite the existence of many sound algorithms for rule induction, the field still enjoys great popularity and, as shown by recent publications (Ishibuchi and Yamamoto, 2005; Cloete and Van Zyl, 2006; Juang et al., 2007; Fernández et al., 2007), offers scope for further improvements.

To find the similarity is very widely useful in real life, because there always have two things with similarity.

For example, search engine - when we need to find something in the internet, people usually use Google, Bing or Baidu these searching engine something, and the key words is an essential condition when we use searching engine, searching engine will list the results it funds. And we can see the keyword is a similarity of these results, maybe these results are not same in detail or in essence, but all the searched results have one common at least, or we can name it as the “overlap”. And they are having more similarity in the top list, their content are connectivity. Duplication check - these problems occur in academic. In the academic, people are attentive to the academic things, and the paper is a very important part of that. When we want to publish our paper in a conference or a journal, at the first step is to check the duplicate contents. To compare how your paper is similar with other papers which are published and can be searched online. We can seem that as the data comparison. Image processing - image processing has a strong connection with data mining; first of all, an image is a dataset technically. A data that can be represented in pixels, and every pixel is a vector which include the color variables. So a whole image is a matrix of pixel data. And a matrix is also a dataset. If two pictures are needed to compare the similarity or distinguish the difference, sometimes we need to judge a picture whether is modified or be changed by Photoshop. It’s very hard to distinguish by visual inspection, and data mining for classification will work. Algorithms will process it in digital way. Many tasks in computer vision involve assigning a label (such as disparity) to every pixel. A common constraint is that the labels should vary smoothly almost everywhere while preserving sharp discontinuities that may exist, e.g., at object boundaries. These tasks are naturally stated in terms of energy minimization. In this thesis, we consider a wide class of energies with various smoothness constraints. And in biomedical field, people usually need to compare the X-ray pictures to pick out the similarity or rather the dissimilarity between two records.

LITERATURE SURVEY

To calculate the similarity of two datasets actually is a necessary and important part in statistics or data mining, even in our life time. The classification algorithms are widely used in different fields, like image processing, medical diagnosis, data filtration, data comparison and so on. And the FURIA is one of the classification algorithms, FURIA is a novel fuzzy rule-based classification method called Fuzzy Unordered Rule Induction Algorithm or FURIA for short, which is a modification and extension of the state-of-the-art rule learner RIPPER (Cohen, 1995). In particular, FURIA learns fuzzy rules instead of conventional rules and unordered rule sets instead of rule lists. Moreover, to deal with uncovered examples, it makes use of an efficient rule stretching method.

There is a similar work in image processing, it’s the Fast-Approximate Energy Minimization via Graph Cuts by Yuri Boykov, Olga Veksler and Ramin Zabin. Their work describes how to get a picture apart by algorithm, we know a picture is made from pixels, but how can I cut a picture, and which is the threshold among pixels is of a question. Pixels have its own code, that means everyone is a set of numbers. We can see it as a vector, and a whole picture is made from pixels. Numerically it is a pixel matrix, a matrix formed by vectors of variables. And now, we also can treat the matrix as a dataset.
Principles of Information Accountability: An eHealth Perspective
[www.igi-global.com/article/principles-of-information-accountability/118221?camid=4v1a](www.igi-global.com/article/principles-of-information-accountability/118221?camid=4v1a)