From Big Data to Smart Data

Teaching Data Mining and Visualization

Antonio Sanchez, Lisa Burnell Ball
Department of Computer Science
Texas Christian University, Fort Worth, Texas, USA
{a.sanchez-aguillar,l.ball}@tcu.edu

Abstract - *The most important part of big data processing is to create knowledge by converting data into smart data. Smart data necessitates both data mining and visualization techniques. We believe it is important to cover these concepts in an undergraduate computer science curriculum and thus have been teaching a data mining and visualization course for several years. In visualization we concentrate on representation and the interaction with data. For data mining we cover several topics, including entropy and Information Gain to select attributes best for classification and prediction. For processing text, statistical algorithms based on Bayes' Theorem are used for filtering and tokenization of the data. Supervised and unsupervised attribute selection is addressed in part by distinguishing between classification and association. The discussion of the kappa metric and confusion tables using cross validation is also covered, as is the use of instance-based learning based on neighborhood correlation for classification and clustering. Finally, for dealing with very large datasets the use of MapReduce and related models are taught.*

Keywords: smart data, data mining, data visualization, undergraduate courses component

1 Introduction: From Information to Knowledge

"Knowledge is experience, the rest is information"
Albert Einstein

Data mining and visualization are essential considering the exabytes of data now available. To go from Big Data to Smart Data, we need to think in terms of Artificial Intelligence (AI) machine learning algorithms to perform the analysis, selection, and definition of patterns, i.e. creating knowledge rather than processing information using traditional database applications. According to McKinsey [9] smart data analytics will be the key to competition, productivity and innovation. Regardless of the domain (e.g. health care, public service, manufacturing), a large shortage of experts is predicted by 2018 with more than 150,000 unfilled openings for those with the required analytical skills [1]. It is time to start doing something and to rigorously teach these skills to undergraduate Computer Science (CS) and Information Technology (IT) students. We have taken this approach and in this paper we discuss the topics we consider relevant to teach including the acquisition of analytical skills using AI machine learning topics.

It is clear that there must be more emphasis on knowledge and choice than on information. Patterns, clusters and classifications are at best the truth but in many cases we only have probable meaningful intelligence and have to make an adequate selection. Data driven analysis is a process by which machine learning algorithms are able to obtain such adequate solutions. It must be emphasized that there is not a single data mining approach, but rather a cadre of techniques that work alone or in combination with each other.

An important element in smart data deals with the proper visualization of the data; in order to obtain the desired response from an audience data must be beautiful, coherent and interactive. Otherwise we lose the audience regardless of the relevance of the data.

The heart of our pedagogical approach assumes a good deal of computer programming, yet we use powerful programming Java libraries, specifically Processing [11] for Data Visualization, Weka [6] for Data Mining and Hadoop [12] for Map/Reduce stream processing. CS and IT students must be aware that discrete programming is the heart of any software development and a desire for it is what makes us love our profession.

2 From Table Data to Data Visualization

"The greatest value of a picture is when it forces us to notice what we never expect to see."
John W. Tukey

We begin our course teaching data visualization, rather than going from the data towards the user. We believe that students must be made aware that any table or graph must have a natural narrative to engage the audience. Understanding the questions that need to be answer is the way to approach a visualization project. After this then go and look for the necessary data. Visualization requires indeed some data mining, but mostly the requirements deal with acquiring, parsing and filtering the data is where emphasis should be placed. Creating meaningful interactive representations is the key to a successful visualization project.

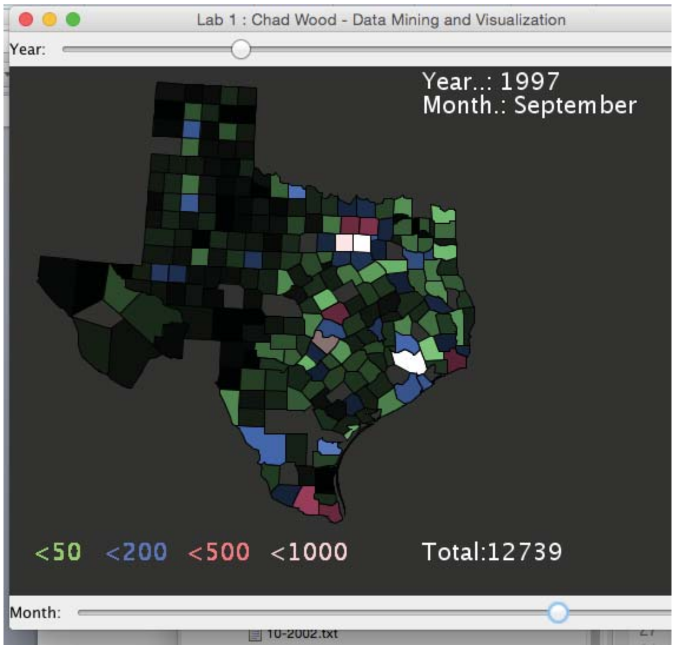
On the subject of representation and interaction we must be careful to direct our interest into representing the dependent variables not in various dimensions; there are only three that we can visualize. Rather try to use in a two dimensional scheme using other means to represent other dimensions. Interaction is to be used to allow the user to dynamically modify the independent variables thus producing the necessary enticement to capture the audiences. Rather than plotting a simple $Y = F(X)$ as a graph or a table of values let the user vary the values of X and let Y change dynamically, as shown in the example presented below.

The example we present is the result of a project assigned to a student [13] and is shown in Fig. 1. In his case rather than presenting tables with public county data for the number of marriages in Texas, he implemented a solution using a SVG map, adding color to the counties with higher values (the dependent variable). The interaction is achieved by modifying the independent variables for month and year thus providing a engaging visualization where patterns can be perceived easily.

For the course we use Ben Fry's processing [3] but not necessarily as a language but rather as a set of Java jar libraries that provide the flexibility to add more jar libraries to enhance any programming project. The benefit of this approach is to have access to the large processing API that reduces the gory details of visual programming as well as file processing and interaction. Moreover, many other capabilities such as data mining implementations are available. Well versed in programming, students can use Eclipse or any other IDE to complete their assignments.

3 Divide and Conquer using Entropy

In Cybernetics entropy has been used as a measurement of diversity and information gain. In his seminal work, W. Ross Quillian approached data mining using entropy to obtain classification trees. To this day C4.5 and C5 are a good approach to do so. The drawback lies in the requirement of preprocessing to create an a priori model. In our course we use Weka [7] libraries to create classification trees; again by using the Weka libraries with Java, students can add more features to their projects, such as nice visualizations as described before. Most importantly, entropy can be used in another crucial task in data mining: attribute classification. This is in the selection of the best feature attributes to be used in the classification, while disregarding those that do not provide useful information gain, reducing the computation time. Entropy as simply defined as $\sum$–$pLog_2p$ is a natural and efficient way to do it. Weka has two classes (InfoGain, GainRatio) that use this metric to evaluate the features or attributes, along with the ranker class for the search method that allows fewer attributes in datasets [8]. Yet the a priori pruning suggested here should be used with care and expertise; this is to say that you must know your data. But then again we are in the realm of knowledge processing and so choices must be made. Generally speaking in a denormalized table the use of entropy can reduce the number of attributes to be considered to a manageable number of less than 10.



Lab 1 : Chad Wood - Data Mining and Visualization

Year:

Year..: 1997
Month.: September

Total:12739

Month:

Fig. 1. Example student visualization project to represent marriage rate by county [13].

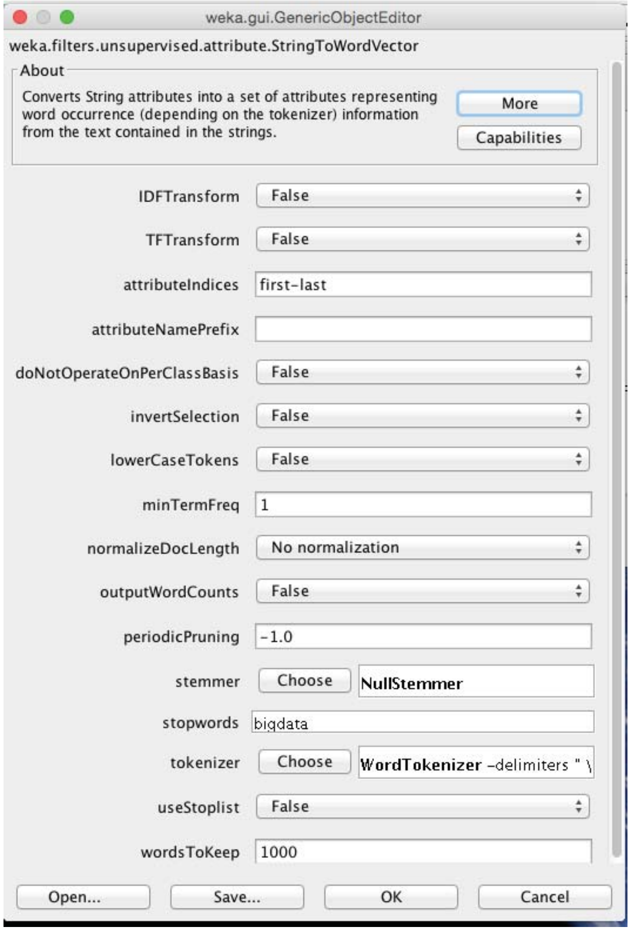
4 Structured Data to Text Classification

Unstructured text processing is an ever increasing application. For example, search engines rely on it. The traditional SQL approach in relational databases prescribes structured, normalized tables. NoSQL alternatives are gaining popularity as alternatives to the relational model. In any case preprocessing text before mining it is an important application to be studied. String tokenization, Stemming and the use of stopword dictionaries are necessary requirements when dealing with text. The computation of text frequencies using IDTF or TFT transforms along with normalization of word frequencies is a detailed and delicate task in the field. Fortunately Weka provides a set of filters that helps in these computations. An example by Hall [6] is shown in Fig. 2.

Classification in this case is done using Bayes' Rule of conditional probability simply defined as

$$p(H|E)=p(E|H)p(H)/p(E) \qquad (1)$$

Although the number of attributes or feature words may be large, the Naïve Bayes approach, which assumes independence among features, is common. Along with this a modified Multinomial approach is also common practice. In any case a Laplace Estimator is used to avoid zero frequencies. We use this approach in the classroom to train text messages for rejection or acceptance.



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

IDFTransform False
TFTransform False
attributeIndices first-last
attributeNamePrefix
doNotOperateOnPerClassBasis False
invertSelection False
lowerCaseTokens False
minTermFreq 1
normalizeDocLength No normalization
outputWordCounts False
periodicPruning -1.0
stemmer NullStemmer
stopwords bigdata
tokenizer WordTokenizer -delimiters " \
useStoplist False
wordsToKeep 1000

Fig. 2. Weka GUI interface showing a filter for string tokenization.

5 Knowledge Validation

The fact that data mining is a selection among alternatives, an important aspect in data mining is that of knowledge validation. Fortunately truth as defined by Saint Thomas is pretty much a metric and not a fixed value, allowing us to discuss in the classroom how good our data mining is. When doing classification using various machine learning algorithms we end up with a single measurement: Success = Right Classifications/Total number of classifications. Provided that we use the normal distribution and the law of large numbers, we can determine confidence level for a classification by solving the probability *p* as suggested by Hall et al. [7]. The important point in this solution is that as the number of cases increase the interval range can be reduced, thus obtaining more precise estimates. Indeed the size of the dataset is ever more important as stated by Halevy, Norvig, and Pereira [5] and Domingos [2]. Added to this a ten-fold cross validation is also a good way to validate a classification. The use of confusion matrices and a Kappa measurement is a standard way to see the performance of a machine learning algorithm since we are reducing the results of a random classificator.

6 Supervised and Unsupervised Classification

Affinity association versus classification is yet another important topic of discussion in class. Mostly because unsupervised attribute selection is the standard approach when confronted with relationships among various activities seen as attributes. The so-called "market basket analysis" occurs when retailers seek to understand the purchase behavior of customers. Weka provides both filtered and unfiltered a priori associators that students can use for Affinity Analysis in contrast to the Prism used in traditional supervised classification.

7 Correlation: the Basis of Big Data Mining

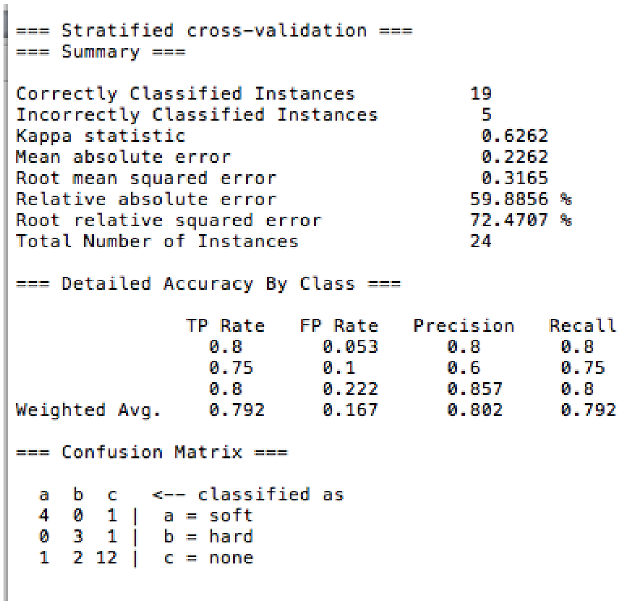
Finally we discuss the most commonly used approach to data mining, instance-based machine learning or memory-based learning based on correlation. In this case instead of performing explicit a priori models, new instances are compared with previous neighborhood instances seen in training, which have been stored in memory. This is the reason for calling this approach lazy learning; in reality it should be called delayed learning. The hypotheses are created dynamically thus allowing data and complexity to grow. Its common practice may be attributed to its ability to adapt to previously unseen data. Instances are compared by a simple numerical distance measurement such as the Euclidean(squared values) or Manhattan (absolute values) using the following simple metric considering all the attributes *(attribute 1 to k)* or features for two instances in question *(i1, i2):*

$$D = SQRT\ [\ (a_1^{i1}-a_1^{i2})^2 + (a_2^{i1}-a_2^{i2})^{2\ +}\ .\ .\ .\ + (a_k^{i1}-a_k^{i2})^2\ ]\quad (2)$$

Using the approach, attribute normalization is required as well as the assignment of maximum values. Examples of this type of classification are the k-nearest neighbor algorithm and kernel machines for limited search. As more data is added memory management becomes an issue when storing all training instances; care should be taken to avoid over fitting to noise in the training set. Clustering is a widely used unsupervised learning approach in which instances are grouped according to a (1) natural center of mass using a distance based measurement or K means, (2) a probabilistic expected maximization, or (3) self organizing. Students need to know how and when to employ clustering. Weka provides useful classes for clustering, examples of which are shown in Fig. 3 [7] and Fig 4 [7].

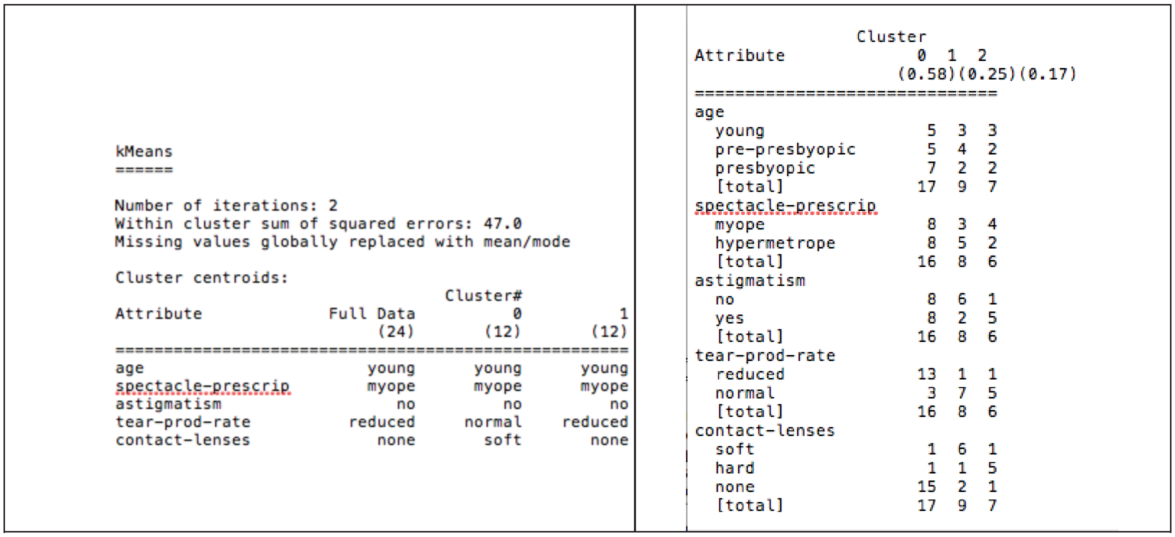
8 Hadoop

A course in Big Data would not be complete without the MAP/REDUCE stream programming model for processing and generating large datasets using a parallel distributed architecture. With data mining an initial set of MAP/REDUCE steps can be perceived as the filtering required to prepare the dataset providing redundant storage.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          19
Incorrectly Classified Instances         5
Kappa statistic                          0.6262
Mean absolute error                      0.2262
Root mean squared error                  0.3165
Relative absolute error                 59.8856 %
Root relative squared error             72.4707 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall
                 0.8       0.053     0.8         0.8
                 0.75      0.1       0.6         0.75
                 0.8       0.222     0.857       0.8
Weighted Avg.    0.792     0.167     0.802       0.792

=== Confusion Matrix ===

  a  b  c   <-- classified as
  4  0  1 |  a = soft
  0  3  1 |  b = hard
  1  2 12 |  c = none
```

Fig. 3. Weka output for instance based classification using a simple dataset for contact lenses.



```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 47.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                       Cluster#
Attribute              Full Data              0             1
                            (24)           (12)          (12)
==============================================================
age                        young          young         young
spectacle-prescrip         myope          myope         myope
astigmatism                   no             no            no
tear-prod-rate           reduced         normal       reduced
contact-lenses              none           soft          none
```

```
                       Cluster
Attribute                  0    1    2
                       (0.58)(0.25)(0.17)
==============================
age
  young                    5    3    3
  pre-presbyopic           5    4    2
  presbyopic               7    2    2
  [total]                 17    9    7
spectacle-prescrip
  myope                    8    3    4
  hypermetrope             8    5    2
  [total]                 16    8    6
astigmatism
  no                       8    6    1
  yes                      8    2    5
  [total]                 16    8    6
tear-prod-rate
  reduced                 13    1    1
  normal                   3    7    5
  [total]                 16    8    6
contact-lenses
  soft                     1    6    1
  hard                     1    1    5
  none                    15    2    1
  [total]                 17    9    7
```

Fig. 4. Weka output for K means and EM clustering with the data used in Fig. 3.

After this first phase data mining, algorithms can be used. The standard software for this approach is Hadoop [12]. For clustering and classification we use mahout [4]. The benefit of using these as Java libraries is that we can combine our programs with other libraries as discussed before. When confronted with massive databases traditional similarity instance-based learning may be substituted with co-occurrence matrix computations that can be obtained using MAP/REDUCE cycles. This is the approach suggested by Owen, Anil & Dunning [10]. Using mahout clustering can also be represented in terms of MAP/REDUCE cycles. Note that Hadoop/mahout can be run both in truly distributed architectures or a single machine thus allowing our Java students to apply their data mining programming skills on their own.

9 Conclusions

"An innovation is a transformation of practice in a community. It is not the same as the invention of a new idea or object. The real work of innovation is in the transformation of practice."
Peter Denning

Indeed research in Big Data and data mining is a hot topic today, however if these breakthroughs are not taught at the undergraduate level the topic does not become an innovation. Just as Peter Denning states, innovation requires new routines to be created and so teaching these concepts becomes an important part of the innovation process. At our institution we teach a course on data mining and visualization open to juniors and seniors; mainly we are interested in savvy programmers that have been exposed to both data structures and database concepts. The course exposes students to current trends, problems, and technologies. Team projects, active learning, and small class size have all contributed to the success of the course. In formal student assessment, students overwhelmingly report the course as challenging and important. We encourage departments to include a course in data mining and visualization in their CS and IT undergraduate programs.

10 References

[1] T. H. Davenport and D. J. Patil, “Data Scientist: The Sexiest Job of the 21st Century,” Harvard Business Review, pp 70-76, October 2012.

[2] P. Domingos, “A Few Useful Things to Know About Machine Learning,” Communications of ACM , vol 55(10), pp 78-87, 2012.

[3] B. Fry, Visualizing Data: Exploring and Explaining Data with the Processing. CA: O’Reilly Media, 2007.

[4] P. Giacomelli. Apache Mahout Cookbook, UK: Packt Publishing Ltd., 2013.

[5] A. Halevy P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data”, IEEE Intelligent Systems, pp. 8-12, March/April 2009.

[6] M. Hall, F. E. Holmes, G. Pfahringer, B. Reutemann, and I. Witten, “The Weka Data Mining Software: An Update”, SIGKDD Explorations, vol 11(1), 2009.

[7] M. Hall, I. Witten, and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques 3rd ed”. MA: Morgan Kaufmann, 2011.

[8] B. Kaluza, Weka How-to. UK: Packt Publishing Ltd, 2013.

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and H. Byers, Big data: The next frontier for innovation, competition, and productivity. NY: McKinsey Global Institute, 2011.

[10] S. Owen, R. Anil, and T. Dunning. Mahout in Action. NY: Manning Publications Co., 2011.

[11] Processing Foundation, *Processing* http://processing.org Accessed March 2, 2015.

[12] T. White, Hadoop: The Definitive Guide 3rd ed. CA: O’Reilly Media, 2011.

[13] C. Wood, “Lab 1 Presentation in CoSc 40023 Data Mining & Visualization”, Fort Worth: TCU, 2014, unpublished.