

Data Mining Learning Bootstrap Through Semantic Thumbnail Analysis

Sebastiano Battiato, Giovanni Maria Farinella , Giovanni Giuffrida, Giuseppe Tribulato

Dipartimento di Matematica ed Informatica – Università di Catania
Viale Andrea Doria, 6 – 95125, Catania
Email: {battiato, gfarinella, ggiuffrida, tribulato}@dmi.unict.it

ABSTRACT

The rapid increase of technological innovations in the mobile phone industry induces the research community to develop new and advanced systems to optimize services offered by mobile phones operators (telcos) to maximize their effectiveness and improve their business. Data mining algorithms can run over data produced by mobile phones usage (e.g. image, video, text and logs files) to discover user's preferences and predict the most likely (to be purchased) offer for each individual customer. One of the main challenges is the reduction of the learning time and cost of these automatic tasks. In this paper we discuss an experiment where a commercial offer is composed by a small picture augmented with a short text describing the offer itself. Each customer's purchase is properly logged with all relevant information. Upon arrival of new items we need to learn who the best customers (prospects) for each item are, that is, the ones most likely to be interested in purchasing that specific item. Such learning activity is time consuming and, in our specific case, is not applicable given the large number of new items arriving every day. Basically, given the current customer base we are not able to learn on all new items. Thus, we need somehow to select among those new items to identify the best candidates. We do so by using a joint analysis between visual features and text to estimate how good each new item could be, that is, whether or not is worth to learn on it. Preliminary results show the effectiveness of the proposed approach to improve classical data mining techniques.

Keywords: Semantic Image Analysis, Data Mining, Content Based Image Retrieval, Learning Bootstrap

1. INTRODUCTION

As mobile phone usage becomes ubiquitous the number of business opportunities for telcos and related operators is growing at a very fast pace. In some countries the mobile phone market is saturated and fierce competition among operators is now based on convincing customers to *switch in* from competitors. This is mostly achieved by offering more attractive rate planes that represent the main customer's decision factor for *churning*.

However, as telco revenues from phone calls reduce, the need for making money from value added services (VAS) becomes imperative. As a matter of fact, mobile data services and applications are fastly increasing in number and complexity [13,14]. Tools such as e-mailing and internet browsing are now available on many devices. M-commerce – the mobile phone counterpart of Internet e-commerce – is also becoming widespread and is proven to be able to produce revenue streams [1]. In some countries, even TV is now available on portable devices. The future of communication is definitively based on: wirelessly, portability and, networking. Mobile phones, better than portable PCs, are able to implement all that.

These additional services need to be optimized in order to maximize their effectiveness. Many variables are available from mobile phone usage logs which are well suited for sophisticated data intelligence analysis such as data and text mining. In particular, for each customer, accurate logs describing his/her device interaction are available. Different data mining models can be exploited in this domain and, recently, telcos are becoming very interested in such techniques as they realize the great benefits they can produce. Academic research in this domain is still in its infancy, many interesting research avenues will be soon opening to researchers. A fairly well established technique is the advertising through SMS and MMS of payable contents such as ring tones, info services (news, weather forecast, sports, etc.), wallpapers, music, videos, and others. The message includes one or more of such offers, in order to try to stimulate the customer

towards a purchase. The number of possible offers to choose from can easily be in the thousands whereas the number of advertising messages (i.e., opportunities) a customer can receive is quite limited (i.e., few a days or even few a week). A large number of messages sent daily (or weekly) to the customer may annoy him/her with the risk of churn. Thus, the problem is how to properly select the best offers for each customer from a wide selection (thousands) while we have very limited number of advertising opportunities (few a week). A careful selection of the number of messages and, mostly, their contents marks the difference between a good and a bad service (i.e., between making or loosing money from that service).

In this paper we describe an experiment to pre-classify new offers based solely on visual and textual features, thus without measuring customer's feedback on those new items, with the intent of reducing the successive learning time. The proposed system combines Data Mining and Computer Vision techniques to extract knowledge from text, logs, and images.

2. THE PROBLEM

Currently we use a clustering model based on customers purchasing record, in other words, we cluster customers based upon their feedback on the various offers we propose over time on their mobile phones (*behavioural analysis*). Each cluster comprises all customers who exhibit similar purchasing trends, that is, who *tend to purchase* a similar set of items. Each item belongs to a defined set of categories, thus, we tend to cluster together customers who denote interests towards same categories.

In order to learn performances of each item we need to carry on a *learning phase*. Typically such a learning phase requires a certain number of *unbiased* exposures of that item to a random set of customers for a period of time. Thus, we measure customers' feedback on receiving that item on their mobile phone. At the end of this phase we can draw conclusions about the *quality* of that item w.r.t. the customer base. Therefore, after we learned on a specific item we can start optimizing its delivery to maximize the likelihood of its purchase (*targeting phase*).

The learning phase can be lengthy and not applicable in some cases. For instance, in our application we receive a large number of new items every day (thus, we need to learn on all of them). However, given the current customer base size and the business policy that restricts on the number of offers sent every day, we do not have enough learning space; in other words, while we are learning on some items a even larger number of new items become available. Under these circumstances we need to decide upon *not learning* on some items with the risk of discarding some good ones.

A related issue is about learning on items that go on sale for specific calendar days, e.g. Halloween. In general, they become available just few days before the event. Thus, we need to sell them before Halloween as they are promptly removed from the listing right after. In this case we need to learn as fast as possible the most likely customers for those items. As already mentioned, a standard learning process may easily take more than few days to complete – that is, Halloween is over and we still did not finish learning.

Thus, we need somehow to select among all those new items arriving daily to identify the most promising ones before starting learning on them. In other words, we need a way to pre-select the best items and then process them through the learning phase.

In our application each commercial offer is composed by a small picture (see figure 1) and an associated short text. For each customer purchase we know the item, the customer id, the purchase timestamp, and the offer mix she received on her mobile phone when she decided to buy. As already mentioned, the problem we are addressing is to *bootstrap* learning (i.e., reducing learning time and cost) for new items by trying to early identify the best ones. The proposed technique is based on picture and text similarities between the new items and the previously purchased ones. We extract a set of visual features from the images and a corresponding set of words. We then rank all new items using a decision tree derived by historical data. Based on such a rank we are now able to drive learning focusing on most promising items.

2.1. Image analysis of visual content

The dataset used in the experiment contains several thousands of RGB thumbnail images with associated advertising short text. All images are encoded by JPEG standard with a high quality setting (i.e. no blocking is evident). Typical resolution size is 200x116. Each image is coupled with a short description message in textual form (maximum 30 chars

long). Figure 1 shows some images belonging to various categories with different scenes, object classes in different semantic context. Our experiments are aimed to make more efficient and effective the bootstrap learning phase combining together some textual features (see below) with a set of visual features. The idea is to characterise in a more robust way the behaviour of the customers capturing their interest respect to the intrinsic visual content present in the images and to use these information for the learning phase. The developed system allows analyzing the visual content of each input image taking into account global and local descriptor by making use of:

Colours – Full histograms and its related statistics respectively in the RGB, HSV and CIE Lab colour spaces;

Filter Response – A complete set of band-pass filters to select both high-frequency details and global appearance. The user also can manually select some specific convolution kernel to highlights some ad-hoc features.

Semantic Analysis – To take into account objects and/or scene context we have built an appearance pixels detector using a Bayesian decision rule for each involved class ([5]). In particular we have implemented a Bayesian classifier mainly based on ([8],[11]).

Considering a specific class, the appearance statistical model is built using histograms with 32 bins per channel in the appearance space (RGB in preliminary experiment). The histograms for *class* and *not_class* are turned into class-conditional probability of *class* and *not_class* appearance as follows:

$$P(\text{appearance} | \text{class}) = \frac{Y(\text{appearance})}{Total_Y} \quad P(\text{appearance} | \neg \text{class}) = \frac{N(\text{appearance})}{Total_N}$$

where $Y(\text{appearance})$ is the number of elements in the *class* histogram bin associated with a specific appearance (an RGB triple in our case), $N(\text{appearance})$ is the number of elements in the $\neg \text{class}$ histogram bin associated with a specific appearance and $Total_Y$ and $Total_N$ are the total number of elements achieved summing all information belonging to the *class* and $\neg \text{class}$ bins respectively. The detector is performed using the likelihood ratio approach:

$$\frac{P(\text{appearance} | \text{class})}{P(\text{appearance} | \neg \text{class})} > \alpha$$

where α is a threshold value. Using the decision rule above, a pixel is considered belonging to a specific *class* if the likelihood ratio is greater than α . The value of α is determined empirically. For each class (e.g. Skin, Sky, Vegetation, etc.) we use an hand-labelled dataset to train (separately for each class) the statistical binary classifier useful at runtime to understand if an image pixel is belonging to a specific class or not (e.g. Skin or not_Skin).

Through such kind of analysis of the input data we are able to manage in a more effective way, the “visual content” obtaining also some insights with respect to the underlying “message” considering its peculiarity (i.e. advertisement, commercial). When you search an image the context and/or the object that you want to observe play a primary role; indeed different objects and related context can be recognized by their appearance. For instance the presence of face in an image can be inferred by “Skin appearance” into pixel domain, outdoor context can be inferred by Sky and/or Vegetation appearance, etc.

All images together with the corresponding feature descriptors are properly stored in a database (SQL based). The system supports the possibility to query the database by example or providing some range search specifying one or more input values. Differently than typical CBIR (Content Based Image Retrieval) solution ([3],[4]), the overall system is in this case integrated with the historical and “commercial” temporal trend of a large dataset of potential customers. Figures 2 shows some snapshots of the related imaging retrieval module. The query results are collected by using the relevance order induced by adding the simple Euclidean distance among each involved descriptor. The actual version does not implement relevance feedback function.

2.2 Text analysis

In this first experiment we used the associated short text simply as a Boolean vector. We collected the set of most common words among all messages in the training set. We then *dummified* our learning space by creating a Boolean variable in our training dataset for each word. Thus, given a phrase, we set to one all Boolean variables corresponding to

the words composing the text, zero otherwise. However, before running such a process we filtered out all *stop words*, then we *stemmed* the remaining ones. Stemming [15] is a basic step in almost every Information Retrieval process that aims to simplify a text document for automatic processing. In particular, it takes care of removing all morphological variants of a single word by reducing a word to its *stem*. For instance, the words “eater” and “eating” are reduced to their (common) root form “eat” without losing their semantic.

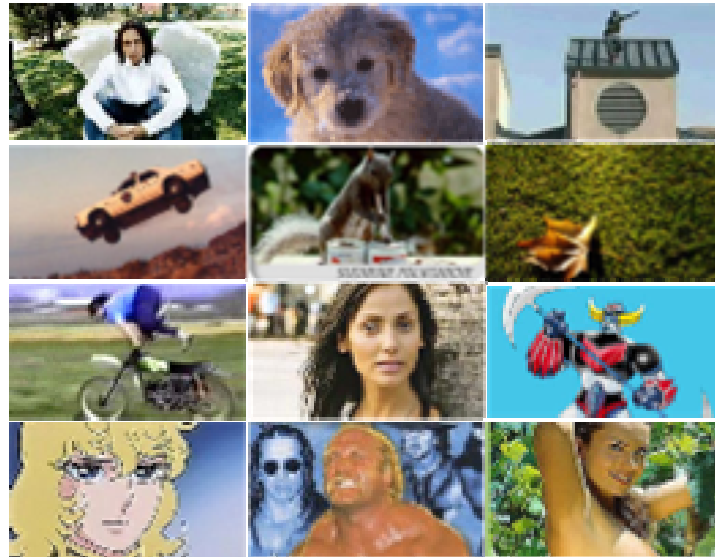


Figure 1 – Some examples of input thumbnail images;

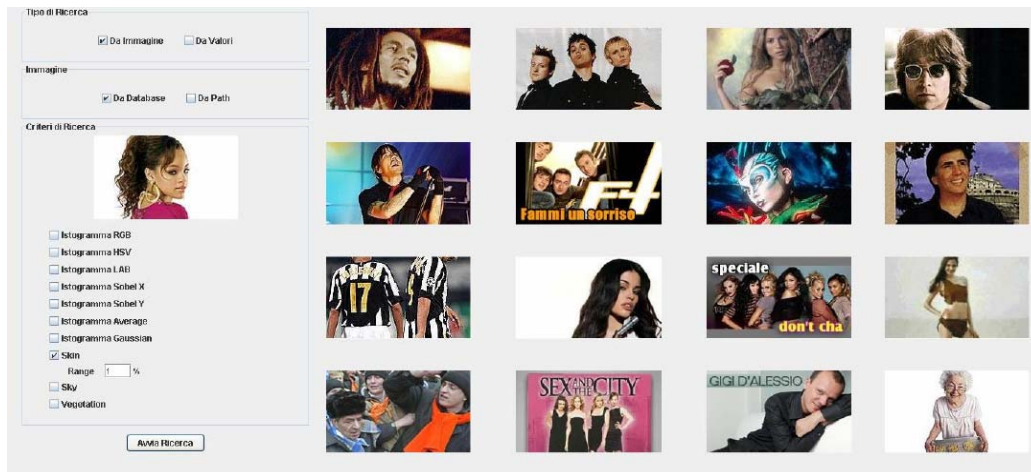


Figure 2 – Query by example, considering only “Skin” feature;

3. COMBINING KNOWLEDGE SOURCES TO BOOTSTRAP LEARNING

The main idea underlying our experiment is to join together image and text analysis on the new items before using them in the cluster analysis. The joint analysis computes for each new item the likelihood of that item to be purchased. Therefore, we can rank all new items from the highest to the lowest probability of being purchased. We then proceed according to the rank in order to include the new items into the clustering model. This allows us to (1) produce more purchases faster and (2) decide wisely about when to abort the learning process if needed – for instance, as already mentioned, when new items arrive and we do not have enough learning space for all of them thus we need to cut on learning somewhere. It is also worth to consider that by combining text with image analysis we exploit a larger set of variables in our model and this makes the model itself more effective.

In this first experiment we used the well known C4.5 [16] algorithm to derive the knowledge model. C4.5 derives a decision tree starting from a classified training set. It accepts both numerical and categorical variables as input whereas the dependent one has to be limited to few possible categories (i.e., classes). C4.5 performs very fast as it follows a divide-and-conquer type of approach. On the arrival of a new item, we perform a generalization of the visual and textual features of the new item with C4.5. The outcome of this process is a rank value associated to each new item. The rank produced is very valuable in those (many circumstances) when a fast learning phase is required. Recall that such ranking is not based on customer purchases but only on static properties (image and text), therefore, it may be applied immediately.

4. EXPERIMENTAL RESULTS

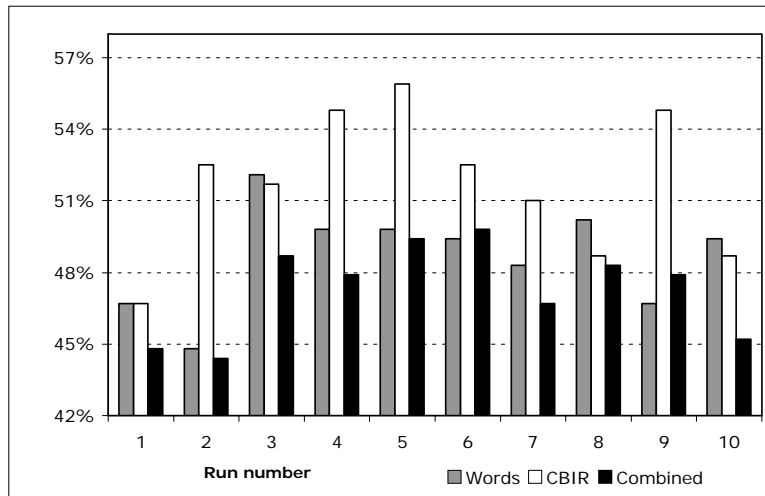
Preliminary results show the effectiveness of the proposed approach to improve classical mining techniques in the field. The overall learning rate of the data mining module, has been sensibly improved both in terms of accuracy and timing; by joining image and text analysis (not fully implemented yet) allows providing a more robust cluster ranking, useful in various context. Also the simple visit of the tree built by C4.5 algorithm allows us to better understand which descriptors can be chosen both in term of text and imaging analysis to guide the learning process.

As already mentioned, in our application we do not have enough space to learn on all new items, therefore we need to discard some or we need to postpone their learning at a better time. By some analysis on the current business trends, if we were going to receive about half of the new offers received daily we could have managed learning for all of them. Based on such a consideration, we decided to reserve some operating margin and therefore to convert our class value into three possible categories: Good, Fair, Bad. We then expect the decision tree generated by C4.5 to produce one of those values during generalization of new items. Thus, after we pre-process all newly received items with the C4.5 we start learning on the ones marked as “Good”. Then, if we still have learning space available we start picking from the “Fair” ones and so on to the “Bad” ones. Therefore, we bootstrap our traditional learning (i.e., measured on user’s feedback) in a sorted fashion starting from the most promising items down to the least performing ones.

Furthermore, we want to test the effectiveness of combining the learning performed on the visual features together with the text ones. To do so, we performed multiple test and for each of them we measured the error rate on the three cases, that is, only visual features (57 independent continuous variables), only text (109 independent Boolean variables), and the combination of the two.

We had a total of 5221 classified tuples. We picked a 5% validation set taken randomly from the input dataset. Given the reduced size of the training set we decided to perform 10 runs of the same test. In each run we were picking a different 5% sample. The following chart shows the complete outcome of our test for the 10 runs of the experiment. In this graph by “Words” we consider the test performed only on the text, “CBIR” the one based solely on the visual features and with “Combined” a combination of the two approaches. Notice that there is some variance in the results of each run which indicates that there are not enough tuples for the number of features. As already mentioned, we mitigate this problem by running the same experiments 10 times on different 5% validation sample. It is important to notice how

the combined approach (the black bar in the chart) performs better than the single ones, that is, the error rate is consistently lower than the two other approaches taken in isolation.



The following table shows the average results computed from the ten runs discussed.

| | Tree nodes number | misclassified test rows | Error % |
|-----------------|-------------------|-------------------------|---------|
| <i>Words</i> | 150,2 | 74,6 | 47,80% |
| <i>CBIR</i> | 397,4 | 78,4 | 50,26% |
| <i>Combined</i> | 485,0 | 72,0 | 46,16% |

Again, notice how the combined approach reduces the overall error rate of the prediction compared to the other two. Consider that if we were going to take randomly one out of three possible classes we were going to have a 66,6% error rate. Thus, the error rate of the combined approach reduces of about twenty percentage points the error due to a random selection. That means we reduced the error by 30%. This represents an interesting results considering that is solely based on static features and not on customers feedback.

The following table shows the *confusion matrix* which shows the error distribution for the three possible classes. The numbers along the diagonal, shown in bold, represent the correct guesses.

| Classified as | | | Real Classes |
|---------------|-----------|-----------|--------------|
| poor | Fair | Good | |
| 54 | 18 | 17 | Poor |
| 35 | 25 | 24 | Fair |
| 15 | 15 | 59 | Good |

It is interesting to notice that if we were going to limit our selection to only the “good” items we were reaching a precision of $59/(17+24+59)=59\%$ that corresponds to an error of 41%. This corresponds to an improvement of $66.6-41 \approx 25$ points improvement, which corresponds to a 37% error reduction.

5. CONCLUSIONS AND FUTURE WORK

In this paper we describe an experiment to address the problem of bootstrap learning for reducing learning time and cost of commercial offers advertised on mobile phones. In the proposed system we combine visual and textual properties of newly arrived items in order to estimate the goodness of such items. We then use such estimates in those circumstances where not enough learning space (or time) is available to learn on all new items. By doing so we are able to focus the learning phase only on the most promising items, thus we bootstrap learning. The features we use in our experiments are not depending on customers' feedback, as they have not even seen yet those items.

We show that under time and space restrictions, instead of picking new learning items randomly we pick them based upon the ranking produced by our system, we reduce the error by over 30%.

Several possible improvements and extensions of the proposed system can be summarized as follows:

- To extend the imaging analysis and retrieval module including some advanced global features useful to capture the visual complexity of the scene in the images ([9],[10], [17]), local features (i.e. textons [12]) which jointly consider shape and texture and geometric context cues of the thumbnails ([6]). Also the possibility to use some ad-hoc classification system for natural scene will be considered ([2],[7], [18], [19], [20]).
- To improve the overall system accuracy by properly choosing a finite set of level for each involved visual feature (e.g., discretization).
- To extend the text analysis module by using ad-hoc techniques able to retrieve some descriptors useful for clustering to be used jointly with other information.
- To measure the effective performances with respect to different pattern classification algorithms such as fwd-forward network [21], SOM [22], Noah [24], and others statistical/probabilistic techniques [23].
- To perform new items classification for each cluster (when a sufficient training set will became available for our experiments): this allows us to increment our clustering-based targeting precision.

Of course, another future extension is the combination of the visual and textual features we used in our pre-learning activity into the clustering modelling we previously developed. This should results in a more precise model as it is based on more variables (and all of them proved to be effective.)

REFERENCES

1. L. Bai, D. C. Chou, D. C. Yen, and B. Lin, "Mobile commerce: its market analyses", International Journal of Mobile Communications, Vol. 3, No. 1, pp. 66-81, 2005;
2. S. Battiato, A. Bosco, A. Castorina, and G. Messina "Automatic Image Enhancement by Content Dependent Exposure Correction", EURASIP Journal on Applied Signal Processing, Vol. 12, No. 12, pp. 1849-1860, 2004;
3. V. Castelli, and D. Lawrence, *Image Databases: Search and Retrieval of Digital Imagery*, Wiley, 2001;
4. Sagarmay Deb, *Multimedia Systems and Content-Based Image Retrieval*, Information Resources Press, 2003;
5. R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley and Sons, 2001;
6. D. Hoiem, A. Efros, and M. Hebert, "Geometric Context from a Single Image", International Conference of Computer Vision (ICCV), pp. 654 - 661 , 2005;
7. F. Naccari, S. Battiato, A. Bruna, A. Capra, and A. Castorina, "Natural Scenes Classification for Color Enhancement", IEEE Transactions on Consumer Electronics, Vol. 51, No. 1, pp.234-239, February 2005;
8. M. J. Jones, and J. M. Rehg, "Statistical Color Models with Application to Skin Detection", International Journal of Computer Vision, Vol. 46, No. 1, pp. 81-96, 2002;
9. A. Oliva, "Gist of a Scene", In Neurobiology of Attention. Eds. L. Itti, G. Rees and J. Tsotsos. Academic Press, Elsevier, pp. 251-256, 2005;
10. A. Oliva, and A. Torralba, "Building the Gist of a Scene: The Role of Global Image Features in Recognition", Visual Perception, Progress in Brain Research, Vol. 155 (in press, 2006);
11. S. L. Phung, A. Bouzerdoum, and D. Chai, "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 27, No. 1, pp. 148-155, 2005;

12. J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation", European Conference on Computer Vision (ECCV), 2006.
13. H. H. Bauer, S. J. Barnes, T. Reichardt, and M. N. Neumann, "Driving Consumer Acceptance of Mobile Marketing: A Theoretical Framework and Empirical Study", Journal of Electronic Commerce Research, Vol 6, N. 3, 2005.
14. E.P. Lim, Y. Wang, K.L. Ong, S.Y. Hwang, "In Search of Knowledge about Mobile Users", ERCIM News, 2003.
15. Hull, D.A., 1996: "Stemming algorithms: a case study for detailed evaluation", Journal of the American Society for Information Science, 47(1), 70-84.
16. J. R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
17. A. Torralba and A. Oliva, "Statistics of natural image categories", Network: computation in neural systems, Vol. 14, 391-412. 2003
18. A. Bosch, A. Zisserman, X. Munoz, "Scene Classification via pLSA", Proceedings of the European Conference on Computer Vision, 2006
19. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, "Discovering object categories in image collections". MIT AI Lab Memo AIM-2005-005, February, 2005.
20. L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, pp. 524-531, 2005.
21. A. K. Jain, J. Mao, and K. M. Mohiuddin. "Artificial neural networks: A tutorial.", IEEE Computer, Vol 29 N.3, pp.56-63, March 1996.
22. T. Kohonen, "Self-organising maps.", Springer-Verlag, Berlin Heidelberg, 1995.
23. A. Jain, P. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions on PAMI 22(1), pp. 4-37, 2000.
24. G. Giuffrida, W. W. Chu, D. M. Hanssens, "NOAH: An Algorithm for Mining Classification Rules from Datasets with Large Attribute Space", Proc. 12th Int'l Conf. on Extending Database (EDBT), Konstanz, Germany, March, 2000.