Intelligent Data Mining For Automatic Face Recognition

Ahmed Ragab, Soumaya Yacout, Mohamed-Salah Ouali

Department of Applied Mathematics and Industrial Engineering
École Polytechnique de Montréal, Canada
ahmed.ragab@polymtl.ca

Abstract: The advancement in computer science and information technology is one of the most important characteristics of the century. One of the important consequences of this advancement is the availability of huge number of automated databases which are waiting to be exploited. This exploitation will lead to knowledge discovery which will help the decision making processes in many fields. In this paper a knowledge discovery, data mining, artificial intelligent technique called Logical Analysis of Data (LAD) is introduced and applied to the well know problem of face recognition. Knowledge discovered in the form of patterns is saved and then used in a machine learning system in order to identify the already learned faces, and to distinguish them from unknown faces. The results show that LAD is promising approach to pattern recognition.

Key words: Data mining, knowledge discovery, artificial intelligence, face recognition, Logical Analysis of Data.

Introduction

The increase in the number of databases in many field create new challenges to researchers in different field. It is reported in (Witten, Frank, & Hall, 2011) that the amount of stored databases doubles every 20 months. It is difficult or even impossible to justify the storage of this amount of data in any quantitative sense. Instead, the important information should be extracted from the data. This operation is called knowledge discovery since there is usually certain amount of useful information that is potentially important in each database and need to be discovered.

Data mining is defined as the process of automatic exploration and extraction of the knowledge from the data (Gorunescu, 2011). The idea is to build computer programs that refine the databases automatically. Among the extracted patterns, some will be trivial and non-interesting, others, on the other hand, will general and can contribute to accurate prediction of future data (Ryoo & Jang, 2009). The patterns discovered must be meaningful and have some advantage in an economic sense. From economical point of view, one of the most important requirements for the patterns is the comprehensiveness (interpretability). Some patterns are comprehensible (also called transparent or interpretable or structural) while some of them are incomprehensible (called black box patterns). From the performance point of view, both of them can make good predictions (Bishop, 2006). The advantage of using comprehensible patterns is their structural representation that can be examined to inform on future events. In other words, they can help the analyst as well the decision maker to explain something about the data in an explicit way (Bores et al., 2000).

Data mining is a topic of increasing interests that involves learning in a practical sense. Researches supported by international agencies, industry and academia are focusing on designing more effective and intelligent data mining techniques (Bozdogan, 2003).

Machine learning provides the technical basis of data mining (machine learning is the technology for mining knowledge from data) (Bishop, 2006). It relies on the availability of data and draw on learning strategies from the area of computational intelligence, statistical pattern classification, and others (Bishop, 2006). The word learning means that these techniques learn from the changes appearing in the data in a way that improves their performance in the future. Thus learning is tied to performance enhancement. Based on this learning process, the learning techniques can be employed to map data into decision model in order to produce predicting output from new data. The decision model is called classifier (Bishop, 2006).

There are two approaches for machine learning, supervised learning and unsupervised learning (Bishop, 2006). In supervised learning, the purpose is to infer the decision model from labeled data. In unsupervised learning, the learning technique is fed with only unlabeled objects (there is no a priori label).

Logical analysis of data (LAD) is a supervised data mining methodology. It was introduced by the group of researchers of RUTCOR at Rutgers University in USA (Bores, et al., 2000). Logical Analysis of Data (LAD) is a combinatorial and optimization based method used in many applications such as oil exploration, detection and prediction of some diseases (Bores, et al., 2000). LAD was introduced to the field of condition based maintenance (CBM) as a new approach for automatic diagnosis of faults in rolling bearings (Mortada, Yacout, & Lakis, 2011). LAD was also able to reproduce human expertise in detecting and analyzing the phenomenon of rogue components in airplanes (Mortada, Carroll, Yacout, & Lakis, 2009). In the airlines industry, LAD was applied to estimate the overbooking level by predicting the show rates of passengers (Dupuis, Gamache, & Pagé, 2012). LAD is applied to develop credit risk rating models for evaluating the credit quality of banks (Hammer, Kogan, & Lejeune, 2012).

One of the advantages of LAD over many data mining techniques is the interpretability (transparency) of its patterns. In other words, LAD can generate patterns that can be easily interpreted and translated into rules which are beneficial to the decision makers (Bores, et al., 2000). LAD is not depending on any statistical analysis; this is another important advantage that makes it capable of dealing with the data that are highly correlated, without the need to satisfy any statistical assumptions.

The main objective of this paper is to apply LAD to the field of face recognition. The aim here is to build a single multi-class decision model that recognizes the images of different objects. This model can effectively deal with multiple changes in facial expression. The paper is organized as follows: The multi-class LAD decision making approach proposed here is presented in the next section. In section 3, the experimental results obtained by using LAD with a known dataset that is employed to train and test different face recognition techniques. Section 4 discusses these results while section 5 concludes the paper.

Logical Analysis of Data

LAD is a combinatorial and optimization method that evolved as an effective classification technique that relies on extracting patterns from binarized data in order to formulate decision rules that classify data into more than one class (Bores, et al., 2000). LAD was used as a Boolean technique to identify the causes of a certain event through investigating a set of factors representing all the possible causes of that event (Crama, Hammer, & Ibaraki, 1988). It is used to extract knowledge from a dataset consisting of observations that can be represented as binary or numerical vectors. Each vector is composed of the values of certain characteristic features). Originally, LAD was used as two-class classification technique (dichotomizer) (Bores, et al., 2000). The observations are classified as either positive Ω^+or negative Ω^- where $\Omega^+$and Ω^- are the sets of positive and negative observations, respectively in the training data setΩ. A specific characteristic of LAD is the extraction of a set of patterns which are the interactions between features for either positive or negative observations in the dataset. Accordingly, LAD can be used as pattern-based classifier of new observations that are not included in the original dataset (Bores, et al., 2000).

Like the conventional two-class LAD, the multi-class LAD decision making approach composed of three steps: data binarization, pattern generation, and theory formation. In what follows, we present the steps of the methodology of LAD which generates an entire set of patterns for a single dichotomy for the two classes. Then we explain how to generate a set of multi-class patterns that can be used to create the decision model in the theory formation step of the multi-class LAD approach.

Data binarization

The binarization procedure in (Mortada, et al., 2011) is presented in this paper, as the first step in LAD methodology. The data binarization step involves the transformation of numerical data to binary data using a binarization technique that transforms each numerical feature into a set of binary attributes. The binarization of a continuous numerical feature A, and the number of binary attributes needed to replace it are dependent on the number of distinct values of A in the training data set. The binarization procedure starts by ranking, in ascending order, all the distinct values of the numerical feature A as follows:

$$u_A^{(1)} < u_A^{(2)} < \cdots < u_A^{(M)} \qquad (M \leq N)$$

Where M is the total number of distinct values of numerical feature A and N is the total number of observations.

Then a cut-point $\alpha_{A,j}$ is introduced between each pair of values that belong to different classes. The cut-point is calculated by averaging the two values as:

$$\alpha_{A,j} = (u_A^{(i)} + u_A^{(i+1)})/2 \quad (1)$$

Where the superscript (i) refers to the order of the distinct value of *A* and *j* refers to a specific feature, $u_A^{(i)} \in \Omega^+$ and $u_A^{(i+1)} \in \Omega^-$ and vice versa. A binary attribute b is then formed from each cut-point. Each cut-point $\alpha_{A,j}$ has a corresponding binary attribute $b_{\alpha_{A,j}}$ with is defined as:

$$b_{\alpha_{A,j}} = \begin{cases} 1 & if \quad u_A \geq \alpha_{A,j} \\ 0 & if \quad u_A < \alpha_{A,j} \end{cases} \quad (2)$$

As a result of this binarization process, the number of binary attributes that make up the binarized training set is equal to the number of cut-points generated for each numerical feature in the training data set.

Pattern generation

Patterns generation is the key building block in LAD decision model. This step is essential in identifying the positive and negative patterns from the binarized dataset of positive and negative observations. The accuracy of LAD decision model depends on the type of generated patterns (Ryoo & Jang, 2009).

i. Definitions and characteristics of Patterns

A positive (negative) *pattern* is defined as an elementary conjunction of some of literals that is true for at least one positive (negative) observation and false for all negative (positive) observations in the training data set (Bores, et al., 2000). A literal is a Boolean variable x or its negation $\bar{x}$. Each binary attribute b_j in the training set can be represented in a pattern by a literal x_j or its negation $\bar{x}_j$, where x_j is used for $b_j = 1$ and $\bar{x}_j$ for $b_j = 0$. The degree d of a pattern indicates the number of literals used in its definition. A pattern is said to *cover* a certain observation if it is true for that particular observation (Bores, et al., 2000). The set of observations covered by the pattern P is denoted as $Cov(P)$. A high degree pattern is more likely to cover small proportion of observations, while pattern with a low degree is more likely to have higher coverage (Ryoo & Jang, 2009). In the testing dataset, misclassified observations are results of generating high degree patterns while unclassified observations are results of low degree patterns (Ryoo & Jang, 2009).

ii. Pattern generation approaches

Patterns are the corner stones in LAD methodology. In the literature, there are three common approaches for pattern generation: enumeration based approaches (Bores, et al., 2000; Hammer, Kogan, Simeone, & Szedmák, 2004), heuristic approaches (Hammer & Bonates, 2006), and Mixed 0-1 Integer and Linear Programming (MILP) based methods (Mortada, et al., 2011; Ryoo & Jang, 2009).

The MILP based approaches proposed in (Ryoo & Jang, 2009) can generate useful patterns that are optimal with respect to various selection preferences (simplicity, selectivity, and evidential (Hammer, et al., 2004)). The procedure for generating one positive (negative) pattern is formulated as an MILP maximization problem. It can generate strong prime patterns which make LAD classifier generalize better on new observations. The experimentations in that paper show that the generated strong prime patterns can reduce the number of unclassified observations (Ryoo & Jang, 2009). The approach can also generate strong spanned patterns and hence the classifier is likely to be robust to noisy observations (reduce the number of misclassified observations) (Ryoo & Jang, 2009). The MILP based method proposed in (Mortada, et al., 2011) is a modified version of the approach introduced in (Mortada, et al., 2011). The modification aims at maximizing the diversity of patterns generated from the same training data set without a significant increase in training time, thus increases the classification power in the two-class problems.

The generation of positive and negative patterns in *two-class LAD* model is extended to *multi-class LAD* decision model. An extension to multi-class applications that involves the modification of the architecture of *LAD* is proposed in (Mortada, Yacout, & Lakis, 2010). The proposed method has the advantage that it generates a less complex decision model which has a better execution time (Mortada, et al., 2010). In that paper, the procedure for pattern generation in multiclass dataset starts by creating empty sets of patterns P_{ij} for each pair of classes (c_i, c_j) where $, j \in \{1,2,..K\}$ $i \neq j$, and K is the total number of classes. The sets P_{ij} are generated through multiple solutions of the MILP based on the single pattern generation algorithm presented in (Mortada, et al., 2011).

Theory formation

The final step in the *LAD* decision model is the theory formation. For the conventional two-class LAD decision model, the generated positive and negative patterns are selected and then used to create a model called the discriminant function that generates a score ranging between -1 and 1. The discriminant function used in (Mortada, et al., 2010) generates a score for each class and therefore the tested observation belongs to the class with the highest score.

Experimental Results

In this section, we explain how multi-class LAD decision model can be used in the field of face recognition. A description of the pre-processing mechanisms used here for extracting features from the images of one of the face dataset in the field is presented. The performance of *multi-class LAD* decision model is compared with other common face recognition techniques.

Japanese Female Facial Expression (JAFFE) database

i. Pre-processing and features extraction

The results presented in this section were all performed on *Japanese Female Facial Expression (JAFFE)* database(Lyons, Akamatsu, Kamachi, Gyoba, & Budynek, 1998). The database contains 203 images of different facial expressions. The images are taken for 10 Japanese female models. Each image is represented as 256×256 pixels. The pre-processing of the images is performed by resizing the images to *100×100* pixels. In order to evaluate the accuracy of the model, we have applied the standard 10-fold cross validation method (Witten, et al., 2011). The *Eigenfaces* and *FisherFaces* are extracted from the training images (Belhumeur, Hespanha, & Kriegman, 1997). The proposed model is compared to some common classification techniques: instant based (IB), Bayesian, support vector machines (SVM), multi-layer perceptron-neural network (MLP-NN) (Witten, et al., 2011). The algorithms for such techniques are implemented in the publicly available Weka software package (Bouckaert et al., 2010).

ii. Performance comparison

The performance comparison between *multi-class LAD* and these classification techniques is shown in Table 1 and Table 2. Table 1 shows that the accuracy is enhanced when the number of *Eigenfaces* increased. In Table 2, the performance is shown for the *FisherFaces*. The objective is to study the impact of changing the number of extracted feature (*Eigenfaces* and *FisherFaces*) on the *LAD* classification accuracy.

Table 1: Eigenfaces with IB, Bayesian, SVM, MLP, and multi-class LAD on JAFFE database

Number of Eigenfaces	1	2	3	5	10	20	40
Minimal Distance classifier (IB)	31.7073	80.4878	85.3659	95.122	97.561	97.561	100
K-Nearst Neighbor (K=5)	41.4634	85.3659	75.6098	85.3659	95.122	95.122	97.561
Multi-Layer Perceptron (MLP)	46.3415	80.4878	85.3659	95.122	97.561	**100**	**100**
SVM with Radial Basis Function	41.4634	73.1707	56.0976	51.2195	34.1463	41.4634	51.2195
Bayesian (Maximum Postorior; MAP)	43.9024	85.3699	85.3659	92.6829	95.122	97.561	97.561
Multi-class LAD	48.7805	85.3659	85.3659	87.8049	95.122	97.561	**100**

Table 2: FisherFaces with IB, Bayesian, SVM, MLP, and multi-class LAD on JAFFE database

Number of Fisherfaces	1	2	3	5	9
Minimal Distance classifier (IB)	70.7317	97.561	**100**	**100**	**100**
K-Nearst Neighbor (K=5)	73.1707	97.561	**100**	**100**	**100**
Multi-Layer Perceptron (MLP)	73.1707	95.122	**100**	**100**	**100**
SVM with Radial Basis Function	68.2927	97.561	**100**	**100**	**100**
Bayesian (Maximum Postorior; MAP)	70.7317	95.122	**100**	**100**	**100**
Multi-class LAD	**80.4878**	**100**	**100**	**100**	**100**

The software cbmLAD is implemented in C++ programming language at École Polytechnique de Montréal, Canada (Salamanca, 2008) is adapted to deal with the special application of LAD to face recognition. The multi-class LAD decision model is trained and tested using the training images of the dataset mentioned above.

Conclusions

This paper aims at exploring an intelligent face recognition technique that employs a database from face recognition literature. *Eigenfaces* and *Fisherfaces* are applied to extract the relevant information from the images which are important for recognition. We described how to propose the *multi-class LAD* classifier as a decision model for the purpose of face recognition. The study shows how *multi-class LAD* and can be utilized and how it might be useful compared to other face recognition techniques. As a final conclusion, the *multi-class LAD* is a promising approach in the field of pattern classification and image processing in particular when it is used with efficient approaches such as *Fisherfaces* that guarantees high discriminative power among the classes. This motivates us to apply *multi-class LAD* as an image classification technique in the context of condition based maintenance in the future.

References

Belhumeur, P.N., Hespanha, J.P., & Kriegman, D.J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19*(7), 711-720.

Bishop, C.M. (2006). *Pattern recognition and machine learning* (Vol. 4): springer New York.

Bores, E., Hammer, P.L., Ibaraki, T., Kogan, A., Mayoraz, E., & Muchnik, I. (2000). An implementation of logical analysis of data. *Knowledge and Data Engineering, IEEE Transactions on, 12*(2), 292-306.

Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2010). WEKA---Experiences with a Java Open-Source Project. *The Journal of Machine Learning Research, 11*, 2533-2541.

Bozdogan, H. (2003). *Statistical data mining and knowledge discovery*: Chapman & Hall/CRC.

Crama, Y., Hammer, P.L., & Ibaraki, T. (1988). Cause-effect relationships and partially defined Boolean functions. *Annals of Operations Research, 16*(1), 299-325.

Dupuis, C., Gamache, M., & Pagé, J.F. (2012). Logical analysis of data for estimating passenger show rates at Air Canada. *Journal of Air Transport Management, 18*(1), 78-81.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12): Springer.

Hammer, P.L., & Bonates, T.O. (2006). Logical analysis of data—an overview: from combinatorial optimization to medical applications. *Annals of Operations Research, 148*(1), 203-225.

Hammer, P.L., Kogan, A., & Lejeune, M.A. (2012). A logical analysis of banks' financial strength ratings. *Expert Systems with Applications*.

Hammer, P.L., Kogan, A., Simeone, B., & Szedmák, S. (2004). Pareto-optimal patterns in logical analysis of data. *Discrete Applied Mathematics, 144*(1), 79-102.

Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998). The Japanese female facial expression (JAFFE) database.

Mortada, M.A., Carroll, T., Yacout, S., & Lakis, A. (2009). Rogue components: their effect and control using logical analysis of data. *Journal of Intelligent Manufacturing*, 1-14.

Mortada, M.A., Yacout, S., & Lakis, A. (2010). Fault Diagnosis of Power Transformers Using Logical Analysis of Data. *APPLICABILITY AND INTERPRETABILITY OF LOGICAL ANALYSIS OF DATA IN CONDITION BASED MAINTENANCE*, 74.

Mortada, M.A., Yacout, S., & Lakis, A. (2011). Diagnosis of rotor bearings using logical analysis of data. *Journal of Quality in Maintenance Engineering, 17*(4), 371-397.

Ryoo, H.S., & Jang, I.Y. (2009). Milp approach to pattern generation in logical analysis of data. *Discrete Applied Mathematics, 157*(4), 749-761.

Salamanca, D. (2008). *The logical analysis of data applied to conditionbased maintenance.* Msc thesis, École Polytechnique, Montréal, Canada.

Witten, I.H., Frank, E., & Hall, M.A. (2011). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.