Association Rule Mining for finding correlations among people

V.B. Nikam[1], Nimai Buch[2], and Yash Botadra[2], B.B. Meshram[3]
[1,2,3]Department of Computer Engineering and Information Technology
Veermata Jijabai Technological Institute
Mumbai, Maharashtra, India

Abstract – *Data mining is the process of extracting interesting, non-trivial, implicit, previously, unknown and potentially useful information or patterns from large information repositories. This paper focuses on Association Rule Mining on large image datasets. ARM is largely applied on datasets containing text, but we shall exploit its capabilities to mine images to get interesting and useful correlations and determine the degree of togetherness among faces in the video. Video processing generates a very large dataset which makes it difficult to analyze it manually. Our research model presented in this paper combines two of the most actively researched areas of computer science: Computer Vision and Data Mining*.

Keywords: Data mining, Association Rule Mining, Computer Vision, Face detection, Face recognition

1 Introduction

Data mining is known as one of the core processes of Knowledge Discovery in Database (KDD) as shown in Fig 1.

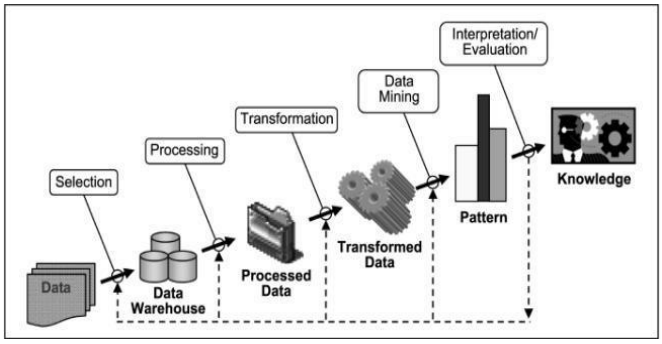

Figure 1 : Knowledge Discovery in Databases

A wide range of industries including retail, finance, health care, manufacturing, transportation and aerospace etc are already using data mining tools and techniques to take advantage of historical data for analysis and predictions for betterment of decision processes. Data mining helps analysts to recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. Association rule mining is one of the most important and well researched techniques of data mining to extract interesting correlations, frequent patterns, associations or casual structures among the sets of items in the transaction databases or other data repositories.

Association rule mining concludes to generate association rules from those large itemsets with the predefined confidence "ρ", say, large itemset $L_k = \{I_1,I_2, \dots ,I_k\}$, where $I_1,I_2,\dots I_n \in I$, the rule can be $\{I_1,I_2, \dots ,I_{k-1}\} \rightarrow \{I_k\}$. Applying confidence, this rule can be determined as interesting or not, and so on. This can be iterated until all the frequent itemsets are over. Though association rules mining is well researched on structured datasets, certainly it can be extended on multimedia datasets also, as there are video and image based applications [2][3] which are in huge demand. In this paper, we have proposed a model for finding togetherness among people in videos using data mining methodologies. In this model we have focused on three major tasks as listed below: 1) Face Detection 2) Face Recognition and Tagging 3) Association Rule Mining on Tagged frames & face. Our model takes a video input since it is a multimedia data type, usually composed of images and audio. As we are processing only on the image portion, we have completely ignored the audio component. We apply association rule mining on the detected faces from the video to find correlations between people in the video.

2 Motivation

Association rule mining has been proved effective for structured datasets. However data mining on the unstructured data sets, especial face image is a hot research area these days. Boosting algorithms reduce the number of computations needed to mine face/non face drastically. The algorithms perform very well even on CPU platform. However, higher resolution images may create a bottleneck for the performance. Some research work optimized the well known computer vision library OpenCV to run not only on Intel platforms, but also on the Cell BE processor. For face detection using Haar-like features and AdaBoost algorithm, their implementation speeds-up computation up to 11x times for 640x480 video resolutions. Ghorayeb et al. proposed a hybrid implementation of AdaBoost for face detection, has not used Haar-like features [4]. Along with the obvious

applications in the fields of biometrics, video surveillance, human computer interaction and image database management, the proposed technology has a wider scope for more interesting applications too. In the current scenario where crimes are increasing and so are the number of criminals, technologies like these can prove to be valuable in finding and recognizing accomplices of criminals. This may also find great relevance in social media and networking, where, on analyzing images, the closeness among friends can be determined. Using this information, appropriate suggestions can be made to users and people. The current technology and the plethora of applications of facial detection, recognition and analysis has been a motivator to perform further research in this speedily growing area of computer science.

3 Literature Review

3.1 Data Mining

Data Mining is the process of discovering interesting knowledge from large amounts of data using various algorithms [5].

Association Rules: An association rule is an implication of the form $X \rightarrow Y$, where $X, Y \in T$, and $X \cap Y = \Phi$. T is the set of objects, also referred to as items. X is called the antecedent and Y is called the consequent of the rule. In general, a set of items, such as the antecedent or the consequent of a rule, is called an itemset.

Support: Each itemset has an associated measure of statistical significance called support. For an itemset $X \in T$, Support(X)=S, if the fraction of records in the database containing X equals S.

Confidence: A rule has a measure of its strength called confidence, defined as the ratio,

$$\frac{support(X \cup Y)}{support(X)} \quad (1)$$

In association rule mining all the generated rules qualify support and confidence greater than minimum support threshold "σ" and minimum confidence thresholds "ρ" respectively. The algorithm mainly has two components,

a. All itemsets that have support above "σ", are called the frequent itemsets. All others are said to be non-frequent or small.

b. For each frequent itemset, all the rules that have minimum confidence, support are generated as,

$$\frac{support(X)}{support(X-Y)} \geq \rho \quad (2)$$

Then the rule, $(X-Y) \rightarrow Y$ is a valid rule for large itemset X and any $Y \in X$.

Algorithms such as: Apriori, AprioriTid, AIS [1] were proposed for mining all association rules. SETM was proposed to mine association rules using relational operations. These algorithms achieved significant improvements over the previous algorithms. Efficient algorithms like Eclat[6], FP-Growth[7], COFI[8], etc [9] for mining association rules are fundamentally different from Apriori algorithm. These algorithms not only reduce the I/O overhead significantly but also have lower CPU overhead for most of the cases.

3.2 Face Detection and Recognition

Face Detection and Recognition from images and videos is emerging as an active research area. Paul Viola and Michael Jones presented a framework for face detection that is capable of processing with high efficiency and accuracy [10]. There are three key contributions, 1. The introduction of a new image representation called the "Integral Image", which allows the features used by the detector to be computed very quickly. 2. A simple and an efficient classifier which is built using the AdaBoost Learning algorithm. 3. A method for combining classifiers in a "cascade" which allows more computation on promising face-like regions rather than background regions.

Definition: Integral Image: The integral image at location (x,y) is the sum of the pixel values above and to the left of (x,y) inclusive.

Eigen pictures (eigenfaces) are used for face recognition. Given the eigenfaces, every face in the database can be represented as a vector of weights. The weights are obtained by projecting the image into eigenface components by a simple inner product operation. A new test image whose identification required is given, is also represented by its vector of weights. The identification of the test image is done by locating the image in the database whose weights are closest, measured by Euclidean distance to the weights of the test image. Eigenfaces approach works well as long as the test image is "similar" to the ensemble of images used in the calculation of eigenfaces. Face recognition systems using the Linear/Fisher Discriminant Analysis as the classifier have also been very successful [11] [12] [13]. LDA training is carried out via scatter matrix analysis. To perform face recognition on a large face dataset but with very few training face images available per class, a holistic face recognition method based on subspace LDA is proposed.

4 Proposed methodology

Our work aims at finding relationship strength between people seen together in a video. This is done by detecting faces present in video frames, recognizing them if seen in earlier part of video, and finally using association rule mining

algorithms to find the association rules for finding togetherness among the people in the video. We perform the image processing tasks using OpenCV, the open source Computer Vision library. The Viola Jones method for facial detection makes use of a cascade of Haar classifiers. These detected faces are then tagged and recognized using Principal Component Analysis (PCA) method, which makes use of eigenfaces and eigenvectors. The recognized faces are a data set of faces in each frame. We finally mine the data set using Association Rule Mining model. The block diagram shown in Fig 2 is a representation of a step-wise execution of our model, taking a raw video as input and generating the correlated group of people using association rules mining as the final output.

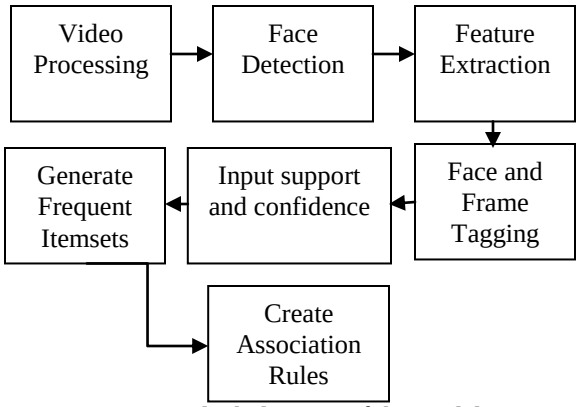

Figure 2: Block diagram of the model

The modules shown in the block diagram are as follows:

4.1 Video processing

The model takes a video as input, on which the face detection and recognition algorithm works. During this step, the video is split into a number of frames on which the Viola Jones algorithm work. The number of frames per second(fps) depends on the quality of the video and the video format (.avi, .wmv, .mp4 etc).

4.2 Face detection

For our model, we make use of the Haar-like cascade of classifiers to detect a face in an image. The cascade of binary classifiers as shown in Fig 3 is applied to check if the portion of the image in the rectangular window qualifies as a face or non face.

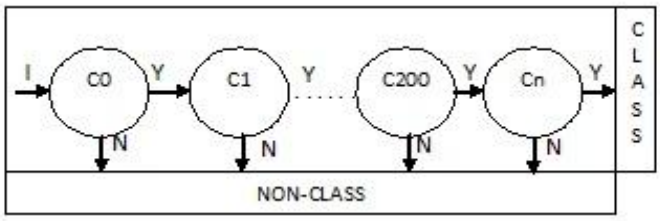

Figure 3: Cascade of binary classifiers

In this image, C_0, C_1, C_2, ..., C_n, is a cascade of "n" classifiers applied over the 24x24 pixel window. Even if the rectangular window does not conform to one of the classifier's requirements, it is not checked any further and is qualified as a non-face.

The basic, weak classifier is based on a very simple visual feature (often referred to as "Haar-like features"). There are four basic Haar features as shown in Fig4.

Haar-like features consist of a class of local features that are calculated by subtracting the sum of a sub-region of the feature from the sum of the remaining region of the feature.

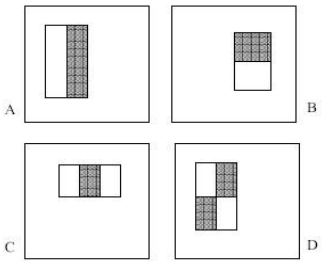

Figure 4 :Basic Haar features

$$f(x,a) = \sum_{i=0}^{n} pb_i \quad (3)$$

If the subtraction of these two regions exceeds the specified threshold value, then the image successfully passes that classifier and a new classifier is applied over it. The process continues for all "n" classifiers and if the image passes successfully through each, it is classified as a face.

4.3 Feature extraction

Feature extraction is the process of applying the Haar feature sub-window of a base size of 24x24 over the image as show in Fig 5. Each of the four feature types are scaled and shifted across all possible combinations.

Figure 5 :Applying Haar features over an image

In a 24x24 pixels sub window, there are ~160,000 possible features to be calculated. These possibilities are reduced to an achievable level using Adaboost techniques [6]. Once the features such as nose, eyes etc are extracted, it becomes relatively easy to detect a face from an image.

4.4 Frame and face tagging

This is the face recognition step, where the detected faces in each frame are compared with a database of faces already detected and tagged. If a match is found, the face is tagged with an existing tag ID else a new tag is assigned to it. Face recognition is done using the Principal Component Analysis (PCA) method which makes use of eigenface and eigenvectors. Fig 6 shows a set of Eigen faces.

Figure 6: Set of Eigen faces

4.5 Support and confidence

The dataset containing faces and the respective frames of the faces is given as input to the association rule mining algorithm. The first step here is to input the Support(s) and Confidence(c) values required to find frequent itemset, and later rule generation for finding co-relations among the faces in the video. Intuitively, a set of faces that appears together in "many" frames is said to be frequent. To define the term "many" we use support "σ" threshold.

4.6 Generate frequent itemsets

The frequent photos are often presented as a collection of if–then rules, called association rules. The form of an association rule is I→j, where I is a set of items and j is an item. The implication of this association rule is that if all of the items in I appear in some basket, then j is "likely" to appear in that basket as well. We formalize the notion of "likely" by defining the confidence of the rule I→j to be the ratio of the support for "I U {j}" to the support for "I". That is, the confidence of the rule is the fraction of the baskets with all of "I" that also contain "j". Using these rules, we can determine the closeness among the people in the video.

Pseudo codes

Detect_And_Recognize_Faces(Video V)

Input: Video
Output: Dataset of Faces in the Video
// This procedure performs operations on videos
// and creates a dataset of faces present in the
// input video.

1) Frames F = Set of frames of the video
2) Dataset D = Φ /*dataset to be generated*/
3) For each imageframe f in F do
   BEGIN:
   a. I = Convert_face-image_to_grayscale(f)
   b. M = intensity_matrix(I)
   c. IM = get_integral_image_matrix(M)
   d. f =get_set_of_faces_detected(IM)
   e. F' = Φ /*face ids in this frame*/
   f. For each face þ in f do:
      BEGIN
      a) fid=Recognize_from_Database(þ)
      b) If þ is not recognized, fid=assign new id to þ
      c) F' =F' U fid
      END
   g. D=D U F'
   END
4) Return D

Generate_Association_Rule(Dataset D, Support S, Confidence C)

Input : Dataset D of Faces, Support S, Confidence C
Output: Togetherness among people
// This procedure performs association rules mining
// operations on distinct tagged faces of each frame of
// video, and finds the co-relations among people in the
// video.

1) FIS = Association_Rule_Mining_Algorithm(D,S)
2) R=Φ /*Set of Rules*/
3) For each frequent item set fis' in FIS do
   BEGIN
   a. For each non empty c in fis' do
      BEGIN
      a) If support(fis')/support(c) ≥ C
         Then R= R U {c → (fis'- c)}
      END
   END
4) Return R

The above pseudo-code describes the general outline of the proposed model, which can be further extended and refined while implementing the same.

5 Conclusion and Future Scope

The KDD process concludes with some knowledge generated from the source data. When finding correlations among the people in a video becomes manually infeasible, association rule mining on video datasets provide results faster. Video processing involves feature extraction and then processing the features for mining. This generates a very huge data set which is practically impossible to process without parallel processing, especially if the video is long and the time required to get output is expected to be very less. However, the same can be achieved with a GPU or cloud like scalable and massively parallel [3][14] processing environments. Our model processes videos to extract correlations among the people in the video. We tag the frames and faces, before we process for co-relation determination. The image indexing can also be extended as a future research direction, which may speed-up the performance of overall processing. The percentage accuracy of the face matching is the further challenge which can be separately addressed in future scope.

6 References

[1]. R. Agrawal, T. Imielinski and A. Swami. "Mining association rules between sets of items in large databases", *International Conference on Management of Data*, Proceedings of ACM SIGMOD, pages 207–216, Washington, DC, May 26-28 1993.

[2]. V.B. Nikam, B.B. Meshram, V.J. Kadam, Image Compression Using Partitioning Around Medoids Clustering Algorithm, InternationalJournal of Computer Science Issues,Vol.8, Issue6, Nov2011, ISSN (Online): 1694-0814.

[3]. V.B. Nikam, Kiran Joshi, B.B. Meshram, "An Approach For System Scalability for Video on Demand", Interface,2011

[4]. Ghorayeb, H., Steux, B., Laurgeau, C., "Boosted algorithms for visual object detection on Graphics Processing Units", Proceedings of the 7th Asian conference on Computer Vision, ACCV'06, Volume Part II, Pages 254-263

[5]. V. B. Nikam, B. B. Meshram, "Scalability Model for Data Mining", ICIMT2010, Dec 28-30, 2010, 978-14244-8882-7/2010, IEEE

[6]. Christian Borgelt, "Efficient Implementations of Apriori and Eclat", Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI), Melbourne, Florida, 2003/11/19

[7]. Christian Borgelt, "An Implementation of the FPgrowth Algorithm", Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, ACM 2005/8/21

[8]. Mohammad El-hajj , Osmar R. Zaïane , "COFI Approach for Mining Frequent Itemsets Revisited",DMKD '04 Published in roceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, Pages 70–75, ISBN:1-58113-908, ACM

[9]. A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules", Proceedings of the VLDB Conference, pages 432– 444, Zurich, Switzerland, September 1995.

[10]. Paula Viola and Michael Jones, "Robust Real Time Face Detection", International Journal of Computer Vision, 57(2),137–154, 2004.

[11]. D. L. Swets, J. Weng, "Discriminant Analysis and Eigenspace Partition Tree for Face and Object Recognition from Views", Proceedings of International Conference on Automatic Face and Gesture Recognition, 1996 pp. 192-197.

[12]. W. Zhao, R. Chellappa, A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition", Proceeedings of International Conference on Automatic Face and Gesture Recognition, 1998 pp. 336-341.

[13]. W. Zhao,R. Chellappa, N. Nandhakumar, "Empirical Performance Analysis of Linear Discriminant Classifiers", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1998 pp. 164-169

[14]. V.B. Nikam, B.B. Meshram, "Scalable Frequent Itemset Mining using Heterogeneous Computing: ParApriori Algorithm", Submitted for Publication