www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19842-19850
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.01

Classification and Analysis of Web Multimedia Data using Principal Component Analysis

Siddu P. Algur[1], Basavaraj A. Goudannavar[2*], Prashant Bhat[3]
[1,2,3] Department of Computer Science,
School of Mathematics and Computing Science,
Rani Channamma University, Belagavi, Karnataka, India
Siddu_p_algur@hotmail.com, agbasu@gmail.com, prashantrcu@gmail.com

Abstract: ***Over the web, the size of multimedia data is increasing in a rapid way. Different types of web multimedia data are used by the web users for different applications. These multimedia data belongs to different categories of domains such as Entertainment, Sports, News and Discussion, Music etc. An automatic classification/prediction of web multimedia data without knowing the content is a challenging and complex research aspect. This paper proposes two approaches to classify web multimedia data viz., classification of web multimedia using dimension reduction technique and classification of multimedia data without reducing the dimensions. To reduce the dimension of the web multimedia metadata, we adopt Principal Component Analysis (PCA) technique to reduce the data dimensions (attributes). The proposed PCA technique involves orthogonal transformation of multimedia metadata values, construction of covariance matrix and computation of Eigen values to reduce the dimensions. The reduced and non-reduced multimedia data are classified separately using DT and KNN classifiers. The classification results of reduced and non-reduced dimensions of multimedia data are analyzed, compared and as a task of KDD.***

Keywords: Principal Component Analysis, Decision Tree and K-Nearest Neighbor classification, Knowledge discovery

1. Introduction

The advances in the digital and network technology have produced multimedia information on the Social media websites such as YouTube, Red Tube, and Face Book etc, automatic organizing of multimedia data into different classes is an emerging trend in the area of web multimedia research. Identifying and organizing a domain specific web multimedia data into different categories using Data mining classification techniques is challenging task. The Classification is a supervised Machine Learning technique which assigns labels or classes to different objects or groups [1] [2]. Many classification models/algorithms and data mining and machine learning tools are developed in recent years. In this work, using KNIME data mining tool [3], the web multimedia metadata dimensions are reduced and classified the category of each web multimedia data.

The Principal component analysis (PCA) is probably the most popular multivariate statistical technique and it is used by almost all scientific disciplines [4]. In many real world problems, reducing dimension is an essential step before any analysis of the data. The general criterion for reducing the dimension is the desire to preserve most of the relevant information of the original data according to some optimality criteria. In pattern recognition and general classification problems, methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) have been extensively used [5][6]. PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the important information from the data table and to express this information as a set of new orthogonal variables called principal components. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in maps [7] [8].

The rest of the paper is organized as follows: The section 2 represents related works on the PCA on web multimedia datasets, section 3 represents proposed methodology, section 4 represents performance evaluation analysis of association rule, and finally section 5 represents conclusion and future work.

2. Related Work

Dimensionality reduction of a feature set is a common preprocessing step used for pattern recognition and classification applications. Principal Component Analysis (PCA) is one of the popular methods used, and can be shown to be optimal using different optimality criteria [7]. In this paper proposed a new method for feature selection named principal feature analysis. The method exploits the structure of the principal components of a set of features to choose the principal features, which retain most of the information, both in the sense of maximum variability of the features in the lower dimensional space and in the sense of minimizing the reconstruction error. The proposed method is applied to two applications, face tracking and content-based image retrieval. The results demonstrate that PFA does have comparable performance to PCA. However, for PCA, all of the original features are needed. This point is the main advantage of PFA over PCA: fewer sensors require or fewer features to compute and the selected features have their original physical meaning. When compared to the optimal features selected in the results show that the PFA features are averagely ranked in the top 5% of all possible combinations.

This paper focuses on using independent component analysis of combined text and image data from web pages. This has potential for search and retrieval applications in order to retrieve more meaningful and context dependent content. It is

demonstrated that using ICA on combined text and image features provides a synergistic effect, i.e., the retrieval classification rates increase if based on multimedia components relative to single media analysis. For this purpose a simple probabilistic supervised classifier which works from unsupervised ICA features is invoked [9].

Application of the Principal Component Analysis method allowed models to predict pregnancy to be built. The basis for modeling was the linear combination of the standardized variables describing the quality of the retrieved oocytes and embryos [10]. Models I and III predicted pregnancy in 61% and 63% of cases, respectively, based on the quality of oocytes. However, correctness of classification for models II, IV, V and VI, which predicted pregnancy based on embryo quality, was higher: 64%, 69%, 67% and 80%, respectively. The best prognostic results for pregnancy were obtained in the gestational carrier group (80%) and in the group with egg donation (69%). Our models demonstrate that good quality oocytes (retrieved from a young, healthy donor) or healthy gestational carriers significantly increase the chances for pregnancy.

Identifying the patterns of large data sets is a key requirement in data mining. A powerful technique for this purpose is the principal component analysis (PCA). PCA-based clustering algorithms are effective when the data sets are found in the same location. In applications where the large data sets are physically far apart, moving huge amounts of data to a single location can become an impractical, or even impossible, task. In this paper, we propose a new algorithm for finding the global PCA of distributed data sets. Our method works directly with the data matrices and has a communications requirement of only O(p2[log2 s]), (i.e., independent of n, the number of observations, which is very large). As compared against the DPCA algorithm [11], our algorithm introduces no local PCA approximation errors. We also consider data updating, and we present a method for computing the PCA for the new extended data sets after new data are added [12].

3. Methodology

In this section we propose an effective methodology to extract the metadata from web multimedia files and classify them based on the extracted metadata by applying data mining and PCA techniques. The main idea and motivation of PCA is to reduce the dimensionality of multimedia metadata dataset by identifying the most significant dimensions. For experimental purpose, out of the total metadata dataset, 60% are used for training and remaining 40% are used for testing. The Decision Tree and KNN classification methods are used to classify web multimedia. The classification results with reduced dimensions and classification results with non-reduced dimensions are compared and analyzed.

The system model of the proposed technique, namely, PCA based classification analysis is depicted in the Fig.1. It consists of the following components:

- Web Multimedia Metadata Extraction and pre-processing
- PCA Technique
  - Orthogonal transformation
  - Covariance matrix
  - Eigenvector and Eigenvalues
  - Dimension reduction
- Classification model
- Classification analysis

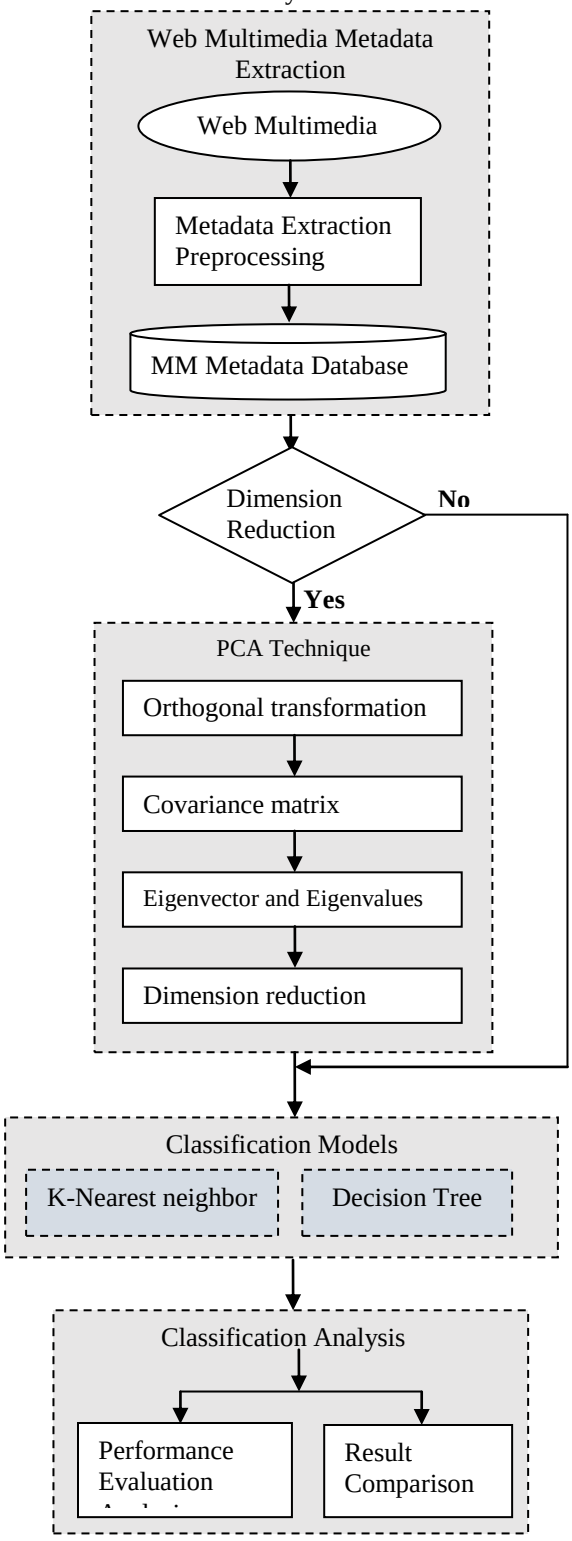

Fig.1. System model of the proposed methodology

The output of each component is the input for the next component. The functionality of each component of the proposed system model is discussed in the following subsections.

3.1 Web Multimedia-Video Metadata Extraction and pre-processing

The metadata of web multimedia-video data are extracted using Mediainfo Extractor tool. Through experimental observation out of 27 attributes 22 attributes found significant for the proposed work. The metadata attributes such as codec

id/info, frame rate mode, color space, scan type and compression mode will be excluded during the experiment because the values of these metadata are constant for each tuple. The remaining twenty two metadata are shown in Table 1. The data are pre-processed for filling missing data with mean of each attribute wherever necessary. The pre-processed and refined multimedia data are stored in a database.

3.2 PCA Technique

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. The central idea and motivation of PCA is to reduce the dimensionality of a point set by identifying the most significant dimensions which are also known as principal components.

3.2.1 Orthogonal Transformation

The PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized in the subspace. Usually, the PCA model will not accept character and special characters. The metadata dimensions "class labels", and "image resolution", are not considered in PCA model, because these metadata dimensions contains special characters as well as characters. The extracted 20 multimedia metadata are stored in the form of CSV data file. The orthogonal transformation multimedia metadata values are shown in Table 2.

3.2.2 Covariance Matrix

A variance-covariance matrix is a square matrix that contains the variances and covariance associated with several variables. The diagonal elements of the matrix contain the variances of the variables and the off-diagonal elements contain the covariance between all possible pairs of variables.

The procedure to calculate covariance matrix is

$$C = \begin{pmatrix} \text{cov}(x,x), \text{cov}(x,z) \ldots \text{cov}(m,n) \\ \text{cov}(y,x), \text{cov}(y,y) \ldots \text{cov}(m,n) \\ \text{cov}(z,x), \text{cov}(z,y) \ldots \text{cov}(m,n) \end{pmatrix} - - - -(1)$$

$$C = \begin{pmatrix} (6.13,\ 194)\ (6.13,586)(6.13,320)(6.13,180) \ldots\ldots\ldots (m,n) \\ (6.13,6.13), (194,194),\ (586,586)(320,320) \ldots\ldots\ldots (m,n) \end{pmatrix}$$

The 494 total video duration is = 3068.4

The average of video duration is = 3068.4/495 = 6.1988

In the multimedia dataset there are 20 columns and 494 rows in the matrix. The covariance matrix value for the multimedia attribute 'video duration' is calculated as follows:

The total duration of 494 'video duration' is 3068 and the average for the same attribute is found 6.1988

To obtain covariance value of 'video duration' the average of 'video duration' will be subtracted by the actual values of 'video duration'

In the same way covariance values for remaining multimedia attributes are calculated and is shown in Table 3.

= 6.1988-6.13 = 0.068

3.2.2 Eigenvector and Eigenvalues

The covariance matrix of the input data and its eigenvectors are used to identifying the directions of maximal variance in the data space. A high value of the eigenvalue indicates a high variance of the data on the corresponding eigenvector. Eigenvectors can be sorted by decreasing eigenvalues, (variance).

The eigenvalues of the matrix: The values of λ which satisfy the characteristic equation of the matrix A, namely

$$\det(A - \lambda I) = 0, \text{ -------- (2)}$$

Where I is the m×n identity covariance matrix, A is reduced PCA dimensional value and λ is multimedia dimensional values.

Form the matrix data (A − λI)

= -0.006x0.068 + -0.324x0.002 + -0.317x-0.001 + -0.352x 0.004 + 0.264x0.00+ 0.192x0.008+ -0.019x0.006+ 0.063x0.001+-0.006x0.068+0.4x0.002+0.384x0.511+.025x-0.005+0.341x0.29+0.269x0.367+0x0.006+0.004x0.012+0.003 x0.01+-0.002x0.014+-0.001x0.014+0.346x0.01

= **0.746**

The eigenvalues and the eigenvector projections necessary to transform each data row from the original space into the new principal components (PCs) space. The overall eigenvalues are shown in Table 4. Usually, only a subset of all PCs is necessary to keep 100% information from the original data set. The more tolerant the closing of information, the higher the dimensionality reduction of the data space.

The procedure to reduce the eigenvalues is

$$C = \frac{\text{total number of eigenvalues}}{\text{number of attributes}} \text{ --------- (3)}$$

$$C = \frac{0.859}{20}$$

C = 0.0042

The 20 multimedia attributes are reduced by using PCA model in to 12 attributes. This scenario is shown in Table 4.

Sl. No.	Attributes	MM_1	MM_2	MM3	MM4	MM5	MM6	MM7	MM8	.....MM_{494}

Table: 1 Multimedia Metadata Values

DOI: 10.18535/ijecs/v6i1.01

1	Video Duration	6.13	2.59	8.49	48.54	2.1	25.51	6.26	4.6	42.54
2	Video Bit rate kbps	194	212	218	234	238	241	245	249	20121
3	Maximum bit rate kbps	586	1596	606	1466	1133	1388	421	574	5403
4	Width Pixels	320	640	320	450	640	640	320	320	1280
5	Height Pixels	180	360	240	360	360	360	240	240	720
6	Display aspect ratio	16.9	16.9	4.3	5.4	16.9	16.9	4.3	4.3	16.9
7	Bits/(Pixel*Frame)	0.134	0.037	0.118	0.058	0.043	0.035	0.127	0.13	0.073
8	Stream size MiB	8.64	4.55	13.8	81.8	223	44.7	11.3	7.32	620
9	Audio Duration	6.13	2.59	8.49	48.54	2.1	25.51	6.26	4.6	42.54
10	Audio Bit rate kbps	72	72	96	96	96	96	96	72	192
11	Maximum bit ratekbps	77.2	76.7	102	103	103	102	105	76.7	201
12	Stream size MiB	3.21	1.54	6.07	33.6	89.9	17.8	4.46	2.12	58.9
13	Image Resolution	320x 180	640x 360	320x 240	450x3 60	640x 360	640x 360	320x 240	320x 240	1280x720
14	Image Height	180	360	240	360	360	360	240	240	720
15	Image Width	320	640	360	450	640	640	320	320	1280
16	Text Page	1	1	1	1	1	1	1	1	1
17	Word Count	15	7	9	24	7	88	25	17	14
18	Character count	88	46	53	143	40	502	144	98	85
19	Line Count	1	1	1	1	1	4	1	1	1
20	Paragraph count	1	1	1	1	1	1	1	1	1
21	Size in kbps	22	25.5	26	22	21.5	23	26	26	25.5
22	class	Sports	News	Sports	News	Ent	News	News	News	Entertainment

Table: 2 Orthogonal transformation for Multimedia data

Sl. No.	**Attributes**	**PCA_0**	**PCA_1**	**PCA_2**	**PCA_3**	**PCA_4**	**PCA_5**	**…..PCA_{494}**
1	Video Duration	3.63	6.65	6.83	3.42	2.11	-7.32	16.84
2	Video Bit rate kbps	981	1563.97	2204.35	1352.95	1616.58	1380.42	-8463.61
3	Maximum bit rate kbps	-10.12	-301.97	6.12	-238.27	-136.61	-214.47	1750.83
4	Width Pixels	-300.66	20.70	-228.73	-169.50	99.22	37.97	-484.05
5	Height Pixels	128.50	122.75	122.76	57.72	-94.36	75.46	-114.34
6	Display aspect ratio	21.36	14.37	50.09	94.95	-73.53	48.99	36.66
7	Bits/(Pixel*Frame)	40.66	86.85	1.99	-59.55	27.38	96.80	-80.30
8	Stream size MiB	33.57	64.51	67.77	-23.99	56.97	-399.21	-9.90
9	Audio Duration	1.00	5.32	-25.14	-2.52	11.19	5.86	9.88
10	Audio Bit rate kbps	6.13	48.79	-28.16	-13.26	3.25	12.76	38.16
11	Maximum bit ratekbps	8.91	5.06	9.63	-47.56	10.53	-13.19	-3.08
12	Stream size MiB	2.56	3.01	1.14	-2.85	0.09	2.02	-0.24
13	Image Height	9.24	1.17	-1.72	-1.67	4.37	4.24	2.60
14	Image Width	-0.75	0.32	-1.15	-0.63	-0.02	0.25	1.90
15	Text Page	-1.17	0.58	0.05	1.20	-0.49	-1.04	1.40
16	Word Count	0.10	0.11	0.03	-0.06	0.07	0.09	-0.05
17	Character count	-0.01	-0.03	-0.03	-0.02	-0.04	0.06	0.29
18	Line Count	0.01	0.22	0.24	-0.15	0.17	0.96	0.05
19	Paragraph count	0.00	0.00	0.00	0.02	0.01	0.00	0.03
20	Size in kbps	0.04	-0.03	0.01	-0.06	-0.02	-0.04	-0.04

International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19842-19850
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.01

Table 3: Covariance matrix of multimedia metadata values

Sl.No.	Attributes	VD	VBR	MBR	WP	HP	DAR	BPF	SM	AD	ABR	MBR	SM	IH	IW	TP	WC	CC	LC	PC	SKbp
1	Video Duration	0.068	0.002	-0.001	-0.004	0	-0.008	0.006	0.001	0.068	0.002	0.001	-0.005	-0.001	-0.005	0	0.004	0.003	-0.002	0	-0.001
2	Video Bit rate kbps	0.002	0.088	0.08	0.081	0.062	0.038	0.001	0.012	0.002	0.098	0.094	0.005	0.081	0.085	-0.001	0.002	0.002	0.003	0	0.001
3	Maximum bit rate kbps	-0.001	0.08	0.085	0.081	0.062	0.038	-0.002	0.016	-0.001	0.094	0.09	0.007	0.079	0.084	-0.001	0.004	0.004	0.005	0	0.003
4	Width Pixels	-0.004	0.081	0.081	0.097	0.071	0.052	-0.007	0.015	-0.004	0.102	0.097	0.008	0.091	0.1	-0.001	0.002	0.002	0.003	0	0
5	Height Pixels	0	0.062	0.062	0.071	0.057	0.028	-0.004	0.008	0	0.079	0.076	0.004	0.07	0.073	-0.001	0.002	0.002	0.003	0	0
6	Display aspect ratio	-0.008	0.038	0.038	0.052	0.028	0.158	-0.012	0.024	-0.008	0.037	0.034	0.009	0.034	0.057	-0.004	0.003	0.004	0.003	0	0.018
7	Bits/(Pixel*Frame)	0.006	0.001	-0.002	-0.007	-0.004	-0.012	0.013	-0.004	0.006	-0.003	-0.003	0	-0.005	-0.008	0	-0.001	-0.001	0	0	-0.003
8	Stream size MiB	0.001	0.012	0.016	0.015	0.008	0.024	-0.004	0.056	0.001	0.016	0.015	-0.001	0.013	0.016	-0.001	0.001	0.001	0	0	0.002
9	Audio Duration	0.068	0.002	-0.001	-0.004	0	-0.008	0.006	0.001	0.068	0.002	0.001	-0.005	-0.001	-0.005	0	0.004	0.003	-0.002	0	-0.001
10	Audio Bit rate kbps	0.002	0.098	0.094	0.102	0.079	0.037	-0.003	0.016	0.002	0.128	0.124	0.005	0.101	0.105	-0.001	0.004	0.003	0.004	0	0.003
11	Maximum bit ratekbps	0.001	0.094	0.09	0.097	0.076	0.034	-0.003	0.015	0.001	0.124	0.12	0.004	0.097	0.101	-0.001	0.004	0.003	0.004	0	0.002
12	Stream size MiB	-0.005	0.005	0.007	0.008	0.004	0.009	0	-0.001	-0.005	0.005	0.004	0.027	0.005	0.008	-0.001	-0.002	-0.002	0	0	0.005
13	Image Height	-0.001	0.081	0.079	0.091	0.07	0.034	-0.005	0.013	-0.001	0.101	0.097	0.005	0.095	0.096	-0.001	0.002	0.001	0.003	0	-0.003
14	Image Width	-0.005	0.085	0.084	0.1	0.073	0.057	-0.008	0.016	-0.005	0.105	0.101	0.008	0.096	0.107	-0.002	0.002	0.001	0.004	0	-0.003
15	Text Page	0	-0.001	-0.001	-0.001	-0.001	-0.004	0	-0.001	0	-0.001	-0.001	-0.001	-0.001	-0.002	0.007	0	0.001	0	0	-0.004
16	Word Count	0.004	0.002	0.004	0.002	0.002	0.003	-0.001	0.001	0.004	0.004	0.004	-0.002	0.002	0.002	0	0.012	0.012	0.008	0	0.003
17	Character count	0.003	0.002	0.004	0.002	0.002	0.004	-0.001	0.001	0.003	0.003	0.003	-0.002	0.001	0.001	0.001	0.012	0.012	0.008	0	0.002
18	Line Count	-0.002	0.003	0.005	0.003	0.003	0.003	0	0	-0.002	0.004	0.004	0	0.003	0.004	0	0.008	0.008	0.009	0	0.002
19	Paragraph count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Size in kbps	-0.001	0.001	0.003	0	0	0.018	-0.003	0.002	-0.001	0.003	0.002	0.005	-0.003	-0.003	-0.004	0.003	0.002	0.002	0	0.052

www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19842-19850
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.01

Table 4: Eigen values of multimedia metadata

.No	Attributes	VD	VBR	MBR	WP	HP	DAR	BPF	SM	AD	ABR	MBR	SM	IH	IW	TP	WC	CC	LC	PC	SKbp	Eigen value	
1	Video Bit rate kbps	-0.006	0.324	0.317	0.352	0.264	0.192	-0.019	0.063	-0.006	0.4	0.384	0.025	0.341	0.367	-0.006	0.012	0.01	0.014	0	0.008	0.746	Filtered in attributes
2	Audio Bit rate kbps	0.388	0.072	0.043	-0.032	0.062	-0.768	0.1	-0.117	0.388	0.141	0.144	-0.069	0.082	-0.047	0.023	0.014	0.003	-0.011	0	-0.137	0.156	
3	Maximum bit rate kbps	0.583	-0.004	-0.02	-0.019	-0.035	0.519	0.007	0.137	0.583	-0.06	-0.068	-0.021	-0.068	-0.015	-0.022	0.047	0.045	-0.006	0	0.109	0.129	
4	Maximum bit ratekbps	-0.007	0.063	0.063	-0.058	0.004	-0.052	0.024	-0.43	-0.007	0.107	0.093	0.123	-0.08	-0.13	-0.065	0.101	0.094	0.079	0	0.844	0.053	
5	Width Pixels	0.046	0.023	-0.083	0.078	0.067	0.21	0.045	-0.868	0.045	-0.086	-0.085	0.057	0.049	0.118	0.031	-0.062	-0.049	-0.02	0	-0.369	0.051	
6	Height Pixels	-0.052	0.015	0.068	-0.084	-0.021	0.068	-0.001	-0.069	-0.053	0.065	0.075	-0.467	-0.073	-0.092	0.057	0.527	0.533	0.372	0	-0.166	0.031	
7	Stream size MiB	0.025	-0.031	0.182	0.071	-0.006	-0.085	0.03	0.079	0.025	-0.148	-0.147	0.811	0.006	0.07	0.039	0.27	0.282	0.263	0	-0.129	0.025	
8	Stream size MiB	0.045	-0.458	-0.291	0.304	0.224	-0.138	-0.442	0.002	0.045	-0.162	-0.186	-0.126	0.325	0.319	-0.041	0.091	0.085	0.082	0	0.19	0.023	
9	Image Height	-0.005	0.333	0.484	0.09	0.066	-0.093	0.167	0.011	-0.003	-0.467	-0.506	-0.266	0.174	0.103	0.013	-0.048	-0.024	0.01	0	0.135	0.014	
10	Image Width	-0.028	-0.046	-0.442	0.074	0.045	0.023	0.793	0.083	-0.029	-0.029	-0.012	-0.015	0.246	0.136	-0.202	0.03	0.022	0.17	0	0.077	0.009	
11	Size in kbps	-0.004	-0.246	0.044	0.115	0.156	0.035	0.245	0.015	-0.004	0.036	0.002	-0.018	-0.02	-0.058	0.904	-0.012	-0.008	-0.084	0	0.092	0.006	
12	Character count	-0.002	0.614	-0.493	-0.203	-0.245	-0.028	-0.234	0.034	-0.001	-0.033	-0.085	0.068	0.249	0.143	0.334	0.095	0.046	-0.017	0	0.06	0.004	
13	Video Duration	-0.012	-0.11	0.095	-0.421	0.276	0.123	-0.05	-0.008	-0.012	-0.015	0.03	0.076	0.674	-0.491	-0.033	0.008	0.004	-0.05	0	-0.048	0.003	Filtered out attributes
14	Audio Duration	-0.019	0.309	-0.272	0.329	0.614	-0.012	-0.033	0.059	-0.019	-0.074	-0.087	0.048	-0.267	-0.382	-0.086	0.143	0.104	-0.259	0	-0.037	0.002	
15	Display aspect ratio	0.021	0.124	-0.077	0.051	0.2	0.012	-0.125	0.029	0.023	0.012	-0.006	-0.022	-0.091	-0.185	0.085	-0.382	-0.24	0.815	0	-0.014	0.001	
16	Bits/(Pixel*Frame)	-0.008	-0.008	0.005	-0.633	0.533	-0.006	0.021	0.029	-0.006	-0.033	0.005	0.013	-0.256	0.491	0.006	0.048	-0.054	0.003	0	0.021	0.001	
17	Text Page	0	0.019	-0.029	-0.039	0.023	-0.009	-0.001	0.007	0.002	-0.195	0.199	0.004	-0.005	0.047	0.006	-0.641	0.706	-0.087	0	0.017	0	
18	Word Count	-0.004	-0.027	-0.002	-0.053	0.02	0.001	0.017	0.002	-0.004	0.692	-0.662	0.007	-0.003	0.003	-0.036	-0.182	0.207	-0.03	0	-0.018	0	
19	Line Count	-0.707	-0.001	0	0.001	-0.001	0	0	0	0.707	0	0	0	0	0	0	0.001	0	-0.001	0	0	0	
20	Paragraph count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19842-19850
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.01

Table 5: PCA reduced dimension values

Sl.No.	Attributes	MM1	MM2	MM3	MM4	MM5	MM6	MM7	.....MM_{494}
1	Video Duration	732	1167	1285	669	698	920	719	977
2	Video Bit rate kbps	76	145	107	71	70	84	76	1680
3	Maximum bit rate kbps	-43	-16	-27	-37	-42	-63	-46	-189
4	Width Pixels	10	17	31	16	17	-17	28	1123
5	Height Pixels	11	-83	-29	18	16	52	12	-231
6	Display aspect ratio	1	64	264	50	30	-13	56	348
7	Bits/(Pixel*Frame)	143	374	488	134	152	171	155	583
8	Stream size MiB	43	-72	101	80	38	178	30	-969
9	Audio Duration	393	853	846	304	407	464	383	9526
10	Audio Bit rate kbps	-135	-454	-371	-58	-126	-77	-125	-2786
11	Maximum bit ratekbps	28	87	89	15	20	46	18	-4525
12	Stream size MiB	-183	-601	-550	-76	-149	-176	-142	9622

Table 6: DT with 12 attributes and DT with 22 attributes

Sl. No	Class Labels	Total Instances	DT with 12 attributes			DT with 22 attributes		
			Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	Sports	85	0.635	0.776	0.698	0.962	0.906	0.933
2	News	100	0.845	0.6	0.702	0.948	0.92	0.934
3	Entertainment	62	0.819	0.952	0.881	0.843	0.952	0.894
	Total	**247**	**0.766**	**0.766**	**0.760**	**0.917**	**0.926**	**0.921**

Table 7: KNN with 12 attributes and KNN with 22 attributes

Sl. No	Class Labels	Total Instances	KNN with 12 attributes			KNN with 22 attributes		
			Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	Sports	85	0.882	0.882	0.706	0.784	0.951	0.906
2	News	100	0.798	0.798	0.83	0.814	0.951	0.97
3	Entertainment	62	0.747	0.747	0.903	0.818	0.891	0.919
	Total	**247**	**0.809**	**0.809**	**0.813**	**0.805**	**0.931**	**0.931**

The Table 4 containing eigenvalues computed from the PCA dimensional values. Each row in the table represents one principal component. The rows are sorted with decreasing eigenvalues, i.e. variance along the corresponding principal axis. The last column in the table contains the component's eigenvalue; a high value indicates a high variance. Each subsequent column (labeled with the name of the selected input values) contains a coefficient representing the influence of the respective input dimension to the principal component.

3.2.4 Dimension reduction

In the dimension reduction phase, PCA-eigenvalue computation will be applied to multimedia dataset to remove the insignificant dimensions. To reduce the dimensions of multimedia data the following steps are used:

- Determine the number of dimensions the input data is projected to.
- The number of target dimensions can either be selected directly or by specifying the minimal amount of information to be preserved.
- If selected directly, the number of dimensions must be lower or equal than the number of input columns.
- Each of the choices for the minimum fraction of information to be preserved corresponds to a possible number of dimensions to reduce to.

- The principal components are created by multiplying the components of each eigenvector by the multimedia PCA dimension values.

The Table 5 shows reduced dimension values of multimedia data

Final Data = Eigen vector x Multimedia PCA dimension values
= 0.746 x 981
= 732

3.3 Classification Model

The proposed work, we adopt two classification models to classify web multimedia video data. The classification accuracy and efficiency will depend on the constructed classification model. This section represents detailed procedure to construct DT and KNN classification model.

3.3.1 Decision Tree Classification Model

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The Decision Tree classification model consist of two major steps i) Attribute selection measures ii) Classification rules. The efficiency of the classification result largely depends on the classification model itself. Hence, construction of robust classification model plays important role in classification. The DT classification model construction for web multimedia-videos is discussed in [13].

3.3.2 K-Nearest Neighbor Classification Model

The k-Nearest Neighbor Classification model is based on learning by analogy, that is, by comparing a given test example with training examples that are similar to it. The training tuples are described by 20 attributes. Each tuples represents a point in a 22-dimensional space. In this way, all of the training tuples are stored in a 20-dimensional pattern space. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training tuples that are closest to the unknown tuples. The KNN classification model construction for web multimedia-videos is discussed in [14].

3.4 Classification Analysis

In this section, performance evaluation measures such as TP, FP, precision, recall and F-Measure will be calculated to measure classification accuracy and efficiency of DT and K-NN classification model. Also the classification accuracy of DT and K-NN will be compared.

4. Experimental Results and Discussions

The multimedia dataset is extracted from the data mining tool and reduced its dimensionality with respect to metadata attributes, by finding a new smaller set of variables using PCA technique which consists of 247 web multimedia metadata instances. The performance of the model is measured in terms of number of correctly classified instances, number of incorrectly classified instances, TP rate, FP rate, precision, recall and F-score.

It is observed from the Decision tree experimental result, out of 247 instances, 185 tuples are correctly classified and 62 tuples are incorrectly classified by the Decision tree classification model. The class labels 'Entertainment' has highest precision and accuracy. Also the falls positive rate of 'News' is very less with respectively. In the 'Entrainment' Class label out of 62 records 59 are correctly classified and 3 were incorrectly classified by DT model. However the false positive rate of class label 'Sports' is high as compare to remaining class label. The overall efficiency of Decision tree classification is found 74.8%.

The K-Nearest Neighbor experimental result, out of 247 instances, 199 tuples is correctly classified and 48 tuples are incorrectly classified by the KNN classification model. The class labels 'Sports' has highest precision and accuracy. Also the false positive rate of 'sports' is very less with respectively. In the 'Entertainment' class label out of 62 records 56 are correctly classified and 6 were incorrectly classified by KNN model. However the false positive rate of class labels 'News' is high as compare to remaining class label. The overall efficiency of KNN classification is found 80.5%.

4.1 Comparative analysis

The quality of the DT and KNN model is represented in terms of comparative analysis and which is represented in Table 6 and 7 as a conclusion of analysis of classification result obtained by the DT and KNN model. Comparative analysis of classification model is found better to classify web multimedia-video data.

The comparative result shows that, DT classification model with 22 attributes, works well as compared to DT classification model with 12 attributes and KNN classification model with 22 attributes, works well as compared to KNN classification model with 12 attributes. However, the results of DT and KNN with 12 attributes are encouraging.

5. Conclusion

This work proposes a novel method for dimensionality reduction of a metadata sets by choosing a subset of the original metadata that contains most of the essential information, using the same criteria as PCA. This paper proposes two approaches to classify web multimedia data viz., classification of web multimedia using dimension reduction technique and classification of multimedia data without reducing the dimensions. To reduce the dimension of the web multimedia metadata, we adopt Principal Component Analysis (PCA) technique to reduce the data dimensions (attributes). The proposed PCA technique involves orthogonal transformation of multimedia metadata values, construction of covariance matrix and computation of Eigen values to reduce the dimensions. The reduced and non-reduced multimedia data are classified separately using DT and KNN classifiers.

References

[1] Ashwini S. Mane, P. M. Kamde, "Video Classification using SVM", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3, July 2013.

[2] S. Syed Shajahaan, S. Shanthi, V. ManoChitra, "Application of Data Mining Techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 11, November 2013.

[3] ChenGang "MediaInfo extractor – A Tool for Media Data Mining", 2011. http://mediaarea.net/en/MediaInfo.
[4] D. S. Hands, "A basic multimedia quality model," IEEE Trans. Multimedia., vol. 6, no. 6, pp. 806–816, IEEE, Transactions On Multimedia, Vol. 6, No. 6, December 2004.
[5] N. Kitawaki, Y. Arayama, and T. Yamada, "Multimedia opinion model based on media interaction of audio-visual communications," MESAQIN 2005, pp. 5–10, June 2005.
[6] Liton Chandra Paul, Abdulla AI Suman, Nahid Sultan "Methodological Analysis of Principal Component Analysis (PCA) Method", IJCEM International Journal of Computational Engineering & Management, Vol. 16 Issue 2, March 2013 ISSN (Online): 2230-7893.
[7] Saporta G, Niang N. Principal component analysis: application to statistical process control. In: Govaert G, ed. Data Analysis. London: John Wiley & Sons; 2009, 1–23. https://cedric.cnam.fr/fichiers/art_1827.pdf
[8] Herve´ Abdi1 and Lynne J. Williams2. "Principal component analysis" WIREs Computational Statistics@2010 John Wiley & Sons, Inc. Volume 2, July/August 2010.
[9] Thomas Kolenda, Lars Kai Hansen, Jan Larsen and Ole Winther. "Independent Component Analysis for Understanding Multimedia Content" Link.
[10] Anna Justyna Milewska, Dorota Jankowska, Dorota Citko, Teresa Więsak2, Brian Acacio3, Robert Milewski1. "The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment Outcome" Studies In Logic, Grammar and Rhetoric 39 (52) 2014. ISBN 978–83–7431–428–2 ISSN 0860-150X.
[11] Y. M. Qu, G. Ostrouchov, N. Samatova, and A. Geist, Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets, Proceedings to the Second SIAM International Conference on Data Mining, April 2002.
[12] Zheng-Jian Bai11, Raymond H. Chan1, and Franklin T. Luk2. "Principal Component Analysis for Distributed Data Sets with Updating" Link.
[13] R. A. Patil, P. G. Ahire, P. D. Patil, Avinash and L. Golande, "Decision Tree Post Processing for Extraction of Actionable Knowledge," International Journal of Engineering and Innovative Technology (IJEIT), Vol. 2, No. 1, 2012, pp. 152-155.
[14] Siddu P. Algur, Basavaraj A. Goudannavar, "Web Multimedia Mining: Metadata Based Classification and Analysis", International Journal of Advanced Research in Computer Science and Software Engineering 5(11), November- 2015, pp. 324-330.
[15] Valdimir Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag,NY, USA, 2000.
[16] Multimedia classification of movie shots using low-level and semantic features. http://articles.ircam.fr/textes/Delezoide05a/index.pdf
[17] Vakkalanka Suresh, C. Krishna Mohan, R. Kumara Swamy, and B. Yegnanarayana, "Content-Based Video Classification Using Support Vector Machines", Springer-Verlag Berlin Heidelberg, LNCS 3316, pp. 726–731 2004.