Feature Ranking Procedure for Automatic Feature Extraction

Sarath krishnan
Department of CSE,
Amrita School of Engineering,Coimbatore
Amrita Vishwa Vidhyapeetham,
Amrita University,
India
sarath.krishnan553@gmail.com

Dr. S. Padmavathi
Department of CSE,
Amrita School of Engineering,Coimbatore
Amrita Vishwa Vidyapeetham,
Amrita University,
India.
s_padmavathi@cb.amrita.edu

***Abstract*—Classifier allows the user to classify between different classes based on the features acquired. The goals and applications of different classifiers are different. As the feature selection is one of the important criteria. In this paper we introduce a method of ranking the features of one class with respect to another and it tells the user that in the training set which feature has higher ranking among the other. So this method tells which feature is insignificant in certain classes and it can be ruled out. The classification can be made so easily as for some cases, certain features creates confusion in the classifier and wrong interpretations are also occurs. In the training set, if a new data is given as input and this method able to tell the user that the features has a variation with respect to training data set and the feature ranking is calculated. This method automatically ranks the feature and feature selection can be made easier. So we can able to interpret from the significant and insignificant features.**

Keywords—classifier;features;ranking;classification;automatic feature selection

I. INTRODUCTION

The general goal of this method as pattern recognition is an important field and it aims to classify objects between different classes based on the features. A Supervised classifier [3] classifies the data from the class information. The features play a vital role in every classifier. Feature selection[2] is one of the important criteria in every pattern recognition problem. The number of feature is too large may leads to curse of dimensionality. QBB-I is one of an efficient feature selection algorithm for very huge data sets and it is considered as a robust algorithm [18].

The machine vision system is an area where pattern recognition has more importance. The system acquires the data in sort of images. There are times when the classifier get confused and gives the wrong output. When some of the features between classes intersect, sometimes it gives a biased output. So in order to avoid such occurrence, ranking of the features between the classes should be done. As Principal Component Analysis is an unsupervised method and there will be no class information [11].

One of the applications considered here is a fruit identification system. It can be used in either grocery shops for computer vision based recognition system. The system can also become helpful for Down syndrome patients. There would be number of challenges to be overcome in this system. As it is a vision based system most important thing we check is the color based recognition. As the fruit ripens the colors may also tend to change. So there would be some sort of misclassification may also leads to false interpretations. So feature selection is one of the important criteria in this system. In the other way if any of the new class arrives and the system automatically finds the ranking between the features and it can be used in automobile industries during assembling as any new material comes in the feature ranking can be used as a lookup table and the measures can also be taken further. Similarly it can also be used for inspection of products in large scale industries.

If red, green and blue is the only factor so it becomes difficult in the basis of classification. So any other feature should also be considered for the ease of identification. In Principal Component Analysis, The eigenvalues and eigenvectors are obtained and it tells you the highest varying principal component as this method gives you which feature has highest between classes.

Figure 1 System Architecture

Section 2 gives the feature extraction. Section 3 discusses the classification techniques. Finally, Section 4 and 5 gives the results and conclusion of this paper.

II. LITERATURE SURVEY

In some of the previous research papers, there are as many algorithms for classification but they are confused on the basis of feature selection. In all machine learning algorithm finding

the significant features is a difficult problem and Chi-Square method is comparatively better in terms of accuracy and performance mainly considered the normalized values [22].

1) PCA

A lot amount of research has been done in fruit sorting system and states that simultaneous sorting of fruits would save time.Polder et al. 2002 used Principal Component analysis with spectral images to grade tomatoes according to its ripeness level[2].Principal Component Analysis is also called as KL transform. It is used to reduce the dimensions.Principal component analysis (PCA) is a technique that is useful for compression and classification of data[11]. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variablesthat nonetheless retains most of the sample's information. The basis functions used in PCA is derived from the input image and it is orthogonal Eigen vectors of the covariance of a data set. This technique is mostly used in previous papers to reduce the number of features. PCA works with Gaussian function and for Non-Gaussian functions Independent Component Analysis can be used[9][11][7]. PCA selects the eigen vectors that have maximum variation and the features they selects were also not known. PCA calculates any number of features and it is represented as eigen values and for dimensionality reduction some values are neglected and it is totally mapped in a new dimension. It selects the high varying components among it but we don't know the distinguishing features from this method. If the feature selection is accurate and then we can give it to any classifier and it classifies with the most appropriate result. As feature selection is the most significant method for any machine learning problems. So the PCA couldn't able to tell the desirable feature in this method as it only gives the mathematical relations. When a new class arrives during training, the entire training procedure has to be repeated from the start. The whole dimension matrix gets disturbed when there is a change in the training set. Distance Classifier

The Euclidean distance classifier used in this paper [22].They implemented a fruit recognition system by taking apple, chickoo, banana and strawberry as dataset images and they had acquired 100 images as training samples. The features they selected for this problem is Mean Red, Mean Green and Mean Blue and the values are also obtained. The system didn't given a higher accuracy on this problem. Apple and strawberry were misclassified as both have a higher amount of red component in it likewise chickoo and bananas were also misclassified because of its color component. Apple has obtained an accuracy of 60%, Banana has obtained an accuracy of 65%,Strawberry has obtained an accuracy of 60% and chickoo has obtained an accuracy of 65%. It gives a poor accuracy in this classifier because as there is lot of misclassification. Feature selection should be done in an efficient manner and the accuracy can be increased as well. Feature selection leads this system to have a poorer accuracy and feature selection is not done by this classifier. For classification in each classifier equal weights has been given to all features and hence misclassification occurs.

B. Bayesian Classifier

It is a fundamental statistical approach for pattern classification and it uses Bayes'theorem to classify the data. Bayesian Classifier is used for classification of fruits in this paper [22].They had acquired 100 samples. The features they selected for this problem is Mean Red, Mean Green and Mean Blue and the values are also obtained. The system didn't given a higher accuracy on this problem. Banana and chickoo showed a higher amount of misclassification as compared to apple and strawberry. Apple has obtained an accuracy of 87.5%, Banana has a least accuracy of 50%, and Strawberry has obtained an accuracy of 83.5% were chickoo showed an accuracy of 64% as well. It gives a higher amount of misclassification has happened in banana using Bayes' classification and the error is not mainly because of the classifier instead feature selection was a problem. If a lot of feature has been taken for this method, Complexity is increased on a higher aspect and it takes a higher time. So selection of features considered a higher preference and feature selection is not been done by this classifier. For classification in each classifier equal weights has been given to all features and hence misclassification occurs.

Feature Selection algorithm QBB-I proved as it is much more efficient and less time consuming as compared to that of QBB and LVF methods for large data sets feature selection[18].

III. PROPOSED SYSTEM

The proposed system is selection of features in a automatic manner. PCA produces a lot amount of features but it does not represent it in the name of the feature as they are mapped in different dimensions as well and we do not know which feature has highest significance in between class problems. When a new training image on the dataset, the whole dimensions gets manipulated and iteration starts from the beginning can cause higher complexity. Samples available are not equal for each class and it varies accordingly. Feature selection is very much important in many cases, the classifier misclassifies because as if two different objects has a same feature highlighted and it may cause misclassification and to avoid this may solve many related problems. Feature selection methods are widely studied and discussed the selection of features which helps the classifier and reduced the complexities [25].

Now a days, A lot amount of features can be extracted from an image as the number of features increases the complexity also gets increased and if it is able to find the some of the distinguishing features it will be an added advantage also. In PCA we can eliminate some of the features but we don't know what all the features we ignored. So in this proposed method we are extracting some of the features and we find which feature has a higher amount of between class

variance and we had ranked it on that basis. So we can able to find that what are the distinguishable features from the above classes and similarly it has been done for every classes. We can also avoid insignificant features as well. This system is considered as an feature selection technique. From the ranking of the features we can give weights for each feature and hence it can also be given to any classifier to get more accurate result during misclassification scenarios. The ranking of the features can be used as a look up table for an user to select the distinguishing features as well. It can also be used as an automatic feature selection tools as well. If a new object came in to the training set the system won't get confused and it calculates the variance as well with the other classes and there is no need of complete disturbance in the vector space. If a variant of the existing class gets in to the system some features will be same and they stands out and hence we can avoid the misclassification and consider it as the same class.

IV. EXPERIMENTAL ANALYSIS

Analysis of my result is explained in this paper and the steps I had done for obtaining results and dataset of different fruit images is obtained from internet. A few set of images is given as training and testing images.

A. Feature Extraction

For each training set of images as it is color image so the RedGreenBlue[RGB] complexions should be checked for differentiation [1]. As every fruit has different color such that finding would be easy in such conditions. Another thing we should consider is the diameter of the images.

1) Mean of Red component

a) For every set of RGB values the mean value is calculated. As mean simply gives the average value of information[14]. Everyimage is comprised of RGB values and the amount of red,green or blue comprises the original color of the image. So each fruit contains a different color and hence the features obtained are mean values of red,green,blue and diameter.

2) Mean of Green component

In each images there would be a green component[14] and all values of the green band has been obtained and hence applied a mean function and hence a value is obtained and it is considered as another feature.

3) Mean of Blue component

The blue component in the image is also calculated by the same method as above. The values are collected and used the mean function to obtain the mean value of the image[14]. Hence any of the dataset is blue in color and the blue value has the maximum value.

In this set of images, the images would be of different colors and it can be understood from the RGB value. It is one of the most important features for such images and hence it should be considered during classification.

4) Diameter value

The diameter of every fruit is calculated from the image and hence every fruit has different diameters. So the diameter of the images is also considered as a very important feature in this set of images.

Similarly the diameter values of every training and testing images are calculated by using these techniques and hence all this 4 values are considered as feature vector space of every image.

B. Methodology

The features of training set images are calculated and it is done for every classes. The values of each features are addedin a class seperately divided by the total number of training for that classes.Similarly the 4 feature vector should be obtained for every classes.

So we will get a consolidated feature vector space for each classes. Hence we get 4 feature vector for each classes and hence the values are simplified. The between class variation should be calculated for every class with respect to another class and it shows the feature wise variation for 2 classes.

Variance

It is the measurement of the spread between the numbers. It is a statistical method[4] to measure the difference in mathematical numerals and it can be used to find the difference as well. There are numerous application by using this method. The formula to calculate variance is,

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right]$$

The ranking of features for every class has been done, so we can easily find out that in this particular classes which feature is more significant and insignificant in those terms a ranking of features has been done. When in a test image an untrained set of image suddenly arrives the classifier won't get confused and it calculates the between class variation of each classes as well as the ranking of features are also done. So the user can identify up to an extend that the related class of an unknown data. In PCA the eigen values and eigen vectors are just known, i.e either D dimension matrix or L dimension matrix can be obtained.But we do not know which feature of a particular two classes is dominant. These data cannot obtaained from PCA method[11]. In this automatic distingushing of feature extraction method tells what are the features has highest significant and what are least significant. Where as in PCA the particular feature was not known and it is considered as an advantage in this method.

TABLE I.

Ranking of features between Nipis and Lemon		
Ranking	Variation of features between classes	Variance
1.	Red Component	1345.5584
2.	Blue Component	643.4127
3.	Diameter	109.661
4.	Green Component	007.7673

TABLE II.

Ranking of features between Lemon and Orange		
Ranking	Variation of features between classes	Variance
1.	Diameter	6123.5556
2.	Green Component	0882.5747
3.	Blue Component	0295.1190
4.	Red Component	0008.0706

TABLE III.

Ranking of features between Nipis and Orange		
Ranking	Variation of features between classes	Variance
1.	Diameter	4594.2959
2.	Blue Component	1810.0438
3.	Red Component	1145.2109
4.	Green Component	0724.7493

If suddenly a new object came in the system won't get collapsed with the previous outputs were PCA dimensions changes. It obtains the features and ranking of features has also been done by this method.

TABLE IV.

Ranking of features between Apple and Nipis		
Ranking	Variation of features between classes	Variance
1.	Diameter	28367.3515
2.	Red Component	01077.9830
3.	Green Component	00277.4598
4.	Blue Component	00190.3020

Similarly the ranking of the features has been done for every other class and hence the results are also obtained.

Conclusion

This method removes the confusions that is created in selecting feature and classifier as well. This method very much reduces the flaws which occur in previous method. It can be further modeled to the testing sessions as well and it improves the classifier accuracies. This method shows higher accuracies as well. In further studies the performance also can be improved by this method.

References

[1] Jain, A and Healey,G, "A multiscale representation including opponent color features for texture recognition", IEEE Transactions on Image Processing vol.7, No.1, pp. 124-128, 1998.

[2] Sarkar, N, and Wolfe, R. R, "Feature extraction techniques for sorting tomatoes by computer vision" Transactions of the ASAE, vol.28, pp.970-979, 1985.

[3] Anderson Rocha, Daniel C. Hauagge, Jacques Wainer, Siome Goldenstein, "Automatic fruit and vegetable classification from images", Computers and Electronics in Agriculture, Vol. 70, pp. 96–104, 2010.

[4] N.A. Campbell, Robust procedures in multivariate analysis I: Robust Covariance estimation, Applied Statistics, 29 (1980), pp. 231-237.

[5] C. Croux and G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika, 87 (2000), pp. 603.

[6] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2008), pp. 1672-1680.

[7] E.J. Candes, X. Li, Y. Ma, and J. Wright, Robust principal component analysis?, Arxiv preprint arXiv:0912.3599, (2009).

[8] M. Basseville, "'Distance measures for signal processing and pattern recognition", Signal Processin9, Vol. 18, No. 4, December 1989, pp. 349 369.Jain, A and Healey,G, "A multiscale representation including opponent color features for texture recognition", IEEE Transactions on Image Processing vol.7, No.1, pp. 124-128, 1998.

[9] R Gonzalez and P Wintz Digital Image Processing Addison-Wesley,1987.

[10] P. H Heinemann, R. Hughes, C . T. Morrow, H.J Sommer, III, R.B. Beelman and P. J Wuest(1994)" Grading of Mushrooms using a machine vision system " , American Society of Agricultural Engineers , vol 37(5) ,pp 1671-1677.

[11] Libin Zhang, Qinghua Yang, Yi Xun, Xiao Chen, Yongxin Ren, Ting Yuan, Yuzhi Tan and Wei Li (2007), "Recognition of greenhouse cucumber fruit using computer vision" New Zealand Journal of Agricultural Research, vol. 50: pp 1293–1298.

[12] Amy L. Tabb, Donald L. Peterson and Johnny Park (2006)," Segmentation of Apple Fruit from Video via Background Modeling " Proceedings of American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting held at Oregon Convention Center Portland, Oregon during 9 - 12 July 2006.

[13] Andrew webb, - Statistical Pattern recognition, second edition, Wiley publication.

[14] S. Arivazhagan, R. Newlin Shebiah, S. Selva Nidhyanandan, Fruit recognition Using Color and Texture, Journal of Emerging Trends in Computing and Information Sciences, VOL. 1, NO. 2, Oct 2010.

[15] Images, D S Guru, Y. H. Sharath, S. Manjunath, "Texture Features and KNN in Classification of Flower", Department of Studies in Computer Science Manasagangotri, University of Mysore.

[16] Woo Chaw Seng, Seyed Hadi Mirisaee, A New Method for Fruits Recognition System. .

[17] SapanNaik and Dr. Bankim Patel, —Usage of Image Processing and Machine Learning Techniques in Agriculture - Fruit Sorting‖, CSI Communications, October 2013.

[18] Prema Nedungadi and Remya, M. Sb, "A scalable feature selection algorithm for large datasets-quick branch & bound iterative (QBB-I)", Smart Innovation, Systems and Technologies, vol. 27, pp. 125-136, 2014.

[19] Jyoti A Kodagali and S Balaji, —Computer Vision and Image Analysis based Techniques for Automatic Characterization of Fruits – a Review‖, International Journal of Computer Applications (0975 – 8887), Volume 50 – No.6, July 2012.

[20] A.C.L. Lino, J. Sanches and I.M.D. Fabbro,‖Image processing techniques for lemons and tomatoes classification,‖ .Bragantia, vol. 67, no. 3,pp. 785-789, 2008.

[21] F.Albregtsen ‖Statistical texture measures computed from gray level concurrences matrices, —Image Processing Laboratory, Department of Informatics, University of Oslo, pp. 1-14, 1995.

[22] L. K. Devi, P. Subathra, Kumar, P. N., V., R., S., D., and B.K., P., "Tweet sentiment classification using an ensemble of machine learning supervised classifiers employing statistical feature selection methods", Proceedings of the Fifth International Conference on Fuzzy and Neuro

Computing (FANCCO2015), Advances in Intelligent Systems and Computing, vol. 415, pp. 1-13, 2015.

[23] F. Lpez-Garca, G. Andreu-Garca, J. Blasco, N. Aleixos and J. M. Valiente,—Automatic detection of skin defects in citrus fruits using a multivariate image,‖ .Computers and electronics in Agriculture,vol. 71, pp. 189-197, 2010.

[24] Seema Paragi, H. Girish,—A Comparative Study of Image Classifiers in a Fruit Recognition System,‖International Journal of Advanced Research in Computer Engineering & Technology,vol. 3, pp. 189-197, 2014.

[25] B.B. Nair, Preetam, M. T. Vamsi, Panicker, V. R., Kumar, G., and Tharanya, A., "A Novel Feature Selection method for Fault Detection and Diagnosis of Control Valves", International Journal of Computer ScienceIssues,2011.