ISSN:0975-9646

A Survey of Data Uncertainty in Face Recognition

Shubhangi G. Khadse
Department of Computer Science and Engineering
G. H. Raisoni Institute of Engineering and Technology for woman, Nagpur, India

***Abstract*—The face images are obtained from different pose, facial expression and illumination, hence the a single image of the face occurred the high uncertainty for the face representation. The images of face should not be the fully accurate representation of the face and it is an observation of the face images. To reducing the uncertainty for representation of the face and improving the accuracy of face recognition, more observation of the same person face images is required in the face recognition. In the real world face recognition system the uncertainty highly occurred because the limited number of available face images of subject and due to this there is high uncertainty is occurred. In this paper we develop the model which is to improve the accuracy in the face recognition by reducing the data uncertainty. The model is to reduce the uncertainty of face images representation by synthesizing the virtual training samples. Here we select the useful training samples that are similar to the test sample from the set of all the original training samples and synthesized virtual training sample.**

***Keywords*—Computer vision; face recognition; machine learning; uncertainty; face images.**

I. INTRODUCTION

Various factors like imperfect samples lead to data uncertainty. There are two ways of representing the uncertainty data. In first way the data is represented by probability distribution rather than deterministic values. In second way data is represented by statistical information for mean and variance. With these two ways the uncertain data is represented and to reduce the uncertain data, the data must be process. These are the methods to process the uncertain data, the Uncertain Data Management, Data Mining and Data Clustering.

Face Recognition being the most attractive biometric technique it is still a challenging task. Various factors like lighting, expression, pose cause the uncertainty. The best way to reduce the uncertainty is to gain more training samples. More training samples reflect more possible variations of the face and less uncertainty of the data and hence the face.

In many practical face recognition applications for the security, such as e-passport, law enhancement and ID-card identification, in the system there is only a single sample per person recorded in the systems. A new technique is described for synthesizing images of the face from the new viewpoints, when only a single 2-D image is available. The synthesized multiple virtual views of a person under different poses and illuminations from a single face image and exploited extended training samples to classify the face images. We noticed that recently proposed sparse representation-based algorithms works good in face recognition and verification. Generally, a smaller norm means a stronger sparsity and conventional sparse representation algorithms are indeed viewed as a problem of minimizing the $l1$-norm of the coefficient vector.

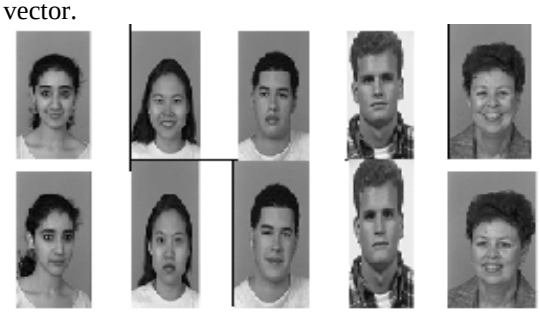

Fig 1. Face images of FERET database

II. RELATED WORK

A. *Robust Classifier*

In propose of this paper designed the problem of a robust classifier this is by minimum the worst case value of a given loss function data overall possible choices of the data in these multi-dimensional intervals. In this paper in detail, the methodology has the application to three specific arising in support vector machines, in logistic regression, loss functions, and in minimum probability machines. In this the resulting problem of convex optimization is amenable to efficient interior-point algorithm. In several practical classification problems, data points are provided approximately, that is their covariates are specified up to given intervals of confidence. For example, when collecting experiments, genomic micro-array data are usually noisy and often a number of replicates of the same experiment are available.

The uncertainty regions can be mathematically specified by a nominal data matrix and a second matrix of the same size containing the corresponding standard errors that bounds inside which every covariate or feature of every data point is known to lay, this leads to a so called interval matrix model for the data. In this paper linear, binary classification problem based on an interval matrix uncertainty model for the data. A robust methodology, where minimize the worst case value of a loss function, over possible realizations of the data within given interval bounds. Shown here how this worst-case loss functions can be upper bounded by a weighted $l1$-norm regularization of the original loss function, explaining the implicit regularization within this approach of robust classification.

In detail three specific choices of a loss function, the primarily the Hinge loss is used in the perspective of soft-margin support vector machines. The secondly loss function is the negative log likelihood function used in

logistic regression. The third loss function is used in the context of minimax probability machines (MPM), which were recently introduced. For each case, we will show that the robust methodology leads to problems that are directly amenable to efficient convex optimization interior-point algorithms. These optimization problems range from linear programming (LP), to second-order cone programming (a generalization of LP which handles *l*2-norm bounds) and constrained maximum entropy for more on interior point methods for convex optimization.

In this paper considered a robust, linear, binary classification problem in which the input data is unidentified but bounded within multi-dimensional intervals. By duality, the interval bounds naturally lead to the presence of weighted *l*1-norms in the constraints imposed on the classifier coefficients these terms induce sparsity of the classifier vector. Thus, robustness and sparsity go together. The convexity, monotonicity and separability properties of the loss function all play an important role of making the robust problem amenable to efficient algorithms for finite-dimensional convex optimization. Our implementation exploits potential regularity, sparsity of the input matrices. [1]

B. Second Order Cone Programming

In this paper, propose a novel second order cone programming formulation which can handle uncertainty in observations for designing robust classifiers. The formulations are resultant for designing regression functions which are robust to uncertainties in the regression setting. Only requiring the existence of second order moments for the proposed formulations are independent of the underlying distribution. These formulations are then focused to the case of missing values in observations for both classification and regression problems and the experiments shows the proposed formulations outperform imputation.

Here considered the problem of binary classification where the labels *y* can take two values, $Y = \{1, -1\}$. This problem was partially addressed, where a second order cone programming (SOCP) formulation was derived to design a robust linear classifier when the uncertainty was described by multivariate normal distributions. The Total Support Vector Classification (TSVC) is another approach, starting from a very similar end up premise, with a non-convex problem with corresponding iterative procedure.

For designing robust binary classifiers developed the proposing a SOCP formulation and for arbitrary distributions having finite mean and covariance and this generalization is done by using a multivariate Chebychev inequality. We also expand this approach to the multicategory case. Next we think the problem of regression with uncertainty in the patterns x, with Chebyshev inequalities two SOCP fromulations are derived, namely Small Residual formulation and Close to Mean formulation, which give linear regression functions robust to the uncertainty in x. As in the classification case the formulations can be interpreted geometrically suggesting various error measures. The projected formulations are then applied to the problem of patterns having missing values both in the case of classification and regression.

Linear Classification by Hyperplanes: - Suppose that we have *n* observations (xi, yi) drawn independently and identically distributed from a distribution over X×Y, where xi is the ith pattern and yi is the corresponding label. To handle uncertainty in the observations the second order cone programming solutions (SOCP) are developed. In this paper they have proposed SOCP formulations for designing robust linear prediction functions which are capable of tackling uncertainty in the patterns both in classification and regression setting. [2]

C. Uncertaindata collection

In propose a number of indirect data collection methodologies have led to the proliferation of uncertain data. These databases are more complex because of the additional challenges of representing the probabilistic information. Here provide a survey of uncertain data mining and management applications and explore the various models utilized for uncertain data representation. The data points may correspond to objects which are only indistinctly specified, and therefore considered uncertain in their representation. The imputation techniques and surveys create data which is uncertain data in nature, this has created a need for uncertain data management algorithms and applications. In uncertain data records, data management is usually represented by probability distributions rather than deterministic values. The main areas of research in the field are as follows:

1) *Modeling of uncertain data:* A key issue is the process of modeling the uncertain data. For that reason, the underlying complexities can be captured while keeping the data useful for database management applications.
2) *Uncertain data management:* One wishes to adapt traditional database management techniques for uncertain data.
3) *Uncertain data mining:* The results of data mining applications are affected by the underlying uncertainty in the data. Hence, it is critical to design data mining techniques that can take such uncertainty into account during the computations.

A number of mining applications have been devised for the case of uncertain data and these type of applications include clustering and classification. The presence of uncertainty can affect the results of data mining applications significantly that can be noticed here.

Clustering Uncertain Data in the presence of uncertainty changes the nature of the underlying clusters, since it affects the distance function computations between different data points. From uncertain data a technique has been proposed in order to find the density-based clusters. The important idea in this approach is to compute uncertain distances effectively between objects which are probabilistically specified. [3]

D. Two-phase Test Sample Representation

In propose a two-phase test sample representation (TPTSR) is the method for face recognition. The firstly the proposed method seeks to represent the test sample as a

linear combination of all the training samples and exploits the representation ability of each training sample to determine M "nearest neighbors" for the test sample. The secondly the test sample as a linear combination of the determined M nearest neighbors and uses the representation result to perform classification. This is helpful to precisely classify the test sample. We will show the prospect explanation of the proposed method.

Transform methods typically use the complete set of training samples to obtain transform axes and then project each test and training sample onto the transform axes to produce a representation for the sample. Then they compute the distance between the representation of the test sample and that of the training sample, and make use of the distance and to classify the test sample use a classifier.

1) *The First Phase of the TPTSR:*-The first phase of the TPTSR uses all of the training samples to represent each and every test sample and develop the representation result to identify the M nearest neighbors of the test sample from the set of the training samples.
2) *The Second Phase of the TPTSR:* - The second phase of the TPTSR generates to represent the test sample as a linear combination of the determined M nearest neighbours and to classify the test sample uses the representation result. [4]

E. Multi-feature Leaning

In propose the Multimedia data are typically represented by multiple features, a new algorithm that is Multi-feature Learning via Hierarchical Regression for multi-media semantics understanding, where two issues are considered. Firstly, labeling large amount of training data is labor intensive; it is meaningful to effectively leverage unlabeled data to facilitate multimedia semantics understanding. Secondly, given that multimedia data can be represented by multiple features, it is valuable to develop an algorithm which combines evidence obtained from different features to infer reliable multimedia semantic concept classifiers.

There are three main strategies to alleviate the tedious work in labeling a large amount of training data for multimedia content analysis. The first strategy is selects the most informative data as the training data which is known as to be labeled active learning. The second one is utilizes the labeled data from another domain which is known as transfer learning. The third one is leverages unlabeled data to infer a more accurate classifier which is known as semi-supervised learning. [5]

III. PROPOSED METHODOLOGY

Here some of the proposed approaches to reduce the data uncertainty.

1) *First approach* - Limited test sample of the subject means limited information what we can do is we can synthesize these limited samples to acquire more possible variations of the face.
2) *Second approach* - Samples obtained in BTTS have both positive and negative effects. So we propose to use the samples from the BTTS which are close to the test samples to represent and classify the test samples.
3) *Third Approach* - Instead of using Conventional Sparse Representation based algorithm which is time consuming go for the l2-norm -based representation algorithm.

In this algorithm solve pose change or misalignment and it requires a test sample to be sparsely represented by a weighted sum of all the training samples. The classification is done by evaluating the demonstration ability on the test sample of each class and by assigning the test sample to the class that has maximum representation ability. Sparse gives that the coefficients of some training samples are equal to zero and the extent of sparsity of the representation coefficients can be measured by the $l1$-norm of the coefficient vector.

IV. CONCLUSION

In this paper, we propose an approach for improving the face recognition accuracy, by reducing the uncertainty. This algorithm produces recognition accuracy in face recognition and produces the proper data analysis. We first exploited the original training samples to synthesize virtual training samples which reflect possible variations of the face and then proposed a scheme of selecting and exploiting useful training samples to represent and classify a test sample. This scheme is helpful for eliminating the improper training samples which have side effect on classification of the test sample, and thus can improve the recognition accuracy.

REFERENCES

[1] L. E. Ghaoui, G. R. G. Lanckriet, and G. Natsoulis, "Robust Classifica- tion with Interval Data," Technical Report UCB/CSD-03-1279, Comput. Sci. Div., Univ. California, Berkeley, Oct. 2003.

[2] P. Shivaswamy, C. Bhattacharyya, and A. Smola, "Second order cone programming approaches for handling missing and uncertain data," J. Mach. Learn. Res., vol. 7, pp. 1283–1314, Jul. 2006.

[3] C. C. Aggarwal and P. S. Yu, "A survey of uncertain data algorithms and applications," IEEE Trans. Knowl. Data Eng., vol. 21, no. 5, pp. 609–623, May 2009.

[4] Y. Xu, D. Zhang, J. Yang, and J. Y. Yang, "A two-phase test sample sparse representation method for use with face recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 9, pp. 1255–1262, Sep. 2011.

[5] Y. Yang, J. K. Song, Z. Huang, Z. G. Ma, N. Sebe, and A. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," IEEE Trans. Multimedia, vol. 15, no. 3, pp. 572–581, Apr. 2013.

[6] B. Qin, Y. N. Xia, S. Prabhakar, and Y. C. Tu, "A rule-based classifica- tion algorithm for uncertain data," in Proc. IEEE 25th Int. Conf. Data Eng., Apr. 2010, pp. 1415–1418

[7] G. Lanckriet, L. Ghaoui, C. Bhattacharyya, and M. Jordan, "A robust minimax approach to classification," J. Mach. Learn. Res., vol. 3, pp. 555–582, Dec. 2002.

[8] P. Shivaswamy, C. Bhattacharyya, and A. Smola, Second order cone programming approaches for handling missing and uncertain data, J. Mach. Learn. Res., vol. 7, pp. 1283–1314, Jul. 2006.

[9] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, Two-dimensional PCA: A new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell., Jan. 2004.