Innovative Constraints Formulation for Max-Margin Hierarchical Clustering

Mr.Akash N. Mhetre
Department of Computer Engineering,RajashreeShahu School of Engineering and Research, JSPM NTC, Pune, INDIA.

Asst.Prof.V.S.Gaikwad
Department of Computer Engineering,RajashreeShahu School of Engineering and Research, JSPM NTC, Pune, INDIA

***Abstract*— Metric learning is nothing but the process of learning distance function over different objects so in that management of large amount of data is little bit difficult process. Manage those data in the form of cluster, nearest neighbors etc. is very important problem that relies on distance function. There are two data types available like linear and nonlinear data. For many types of data, linear model is not very useful but most of metric learning methods assumes linear model of distance. In the recent nonlinear data demonstrated potential power of non-Mahalanobis distance function, particularly tree-based functions. This leads to a more robust algorithm in Metric learning. In that algorithm some technique like k-nearest neighborclassification, large-scale image retrieval and semi supervised clustering problems, then we find that our algorithm yields results comparable to the state-of-the-art. After that compare our method to a number of state-of-the-art benchmarks on k-nearest neighbour classification, large-scale image retrieval and semi supervised clustering problems. A novel tree-based non-linear metric learning method can have information from both constrained and unconstrained points. Combining the output of many of the resulting semi-random weak hierarchy metrics and by introducing randomness during hierarchy training, can obtain a powerful and robust nonlinear metric model.**

Keywords - ***Clustering, data mining, image retrieval, metric learning, non-linear metrics, Constrained and Unconstrained point.***

I. INTRODUCTION

Process of learning distance function over different objects is known as metric learning. In data mining this is very important problem. In data mining as various processes like nearest neighbors, clustering etc. relies on distance function.
Maximum number of metric learning methods assumes linear model of distance. But for many types of data, linear model is not very useful. A wide range of methods have been proposed to address this learning problem, but the field has traditionally been dominated by algorithms that assume a linear model of distance, particularly Mahalanobis metrics [2]. Linear methods have primarily benefited from two advantages. First, they are generally easier to optimize, allowing for faster learning and in many cases a globally optimal solution to the proposed problem[6] Second, they allow the original data simply projected into the novel metric space, meaning the metric can be used in conjunction with different methods that operate only on an explicit feature representation.
This methodology provides two significant contributions: first, unlike previous tree-based nonlinear metrics, it is semi-supervised, and can incorporate information from both constrained and unconstrained points into the learning algorithm. This is mainbenefit in various problem settings, mainly when scaling to longer datasets where only a smallquantity of the full pairwise constraint set can realistically be collected or used in training.Second, the iterative, hierarchical nature of the training process allows us to relax the constraint satisfaction problem. Relativelythan attempting to satisfyaccessible constraint concurrently, at each hierarchy node enhance suitable constraint subset to focus on, leaving others to be addressed lower in the tree. By selecting constraints in this way, we can avoid situations where attempting to satisfy incoherent constraints [11], and thereby better model hierarchical data structures.

II. RELATED WORK

D. M. Johnson, C. Xiong and J. J. Corso proposed obtains more powerful metric model with the help of iterative hierarchical variant of semi supervised max-margin clustering [1]. A. Bellet, A. Habrard, and M. Sebban proposed that recent trends and extensions, such as semi-supervised metric learning, metric learning for histogram data and the derivation of Generalization guarantees, are also covered [2].R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul proposed that developing a new framework of kernelizingMahalanobis distance learners. The new KPCA trick framework offers several practical advantages over the classical kernel trick framework [3].C.Shen, J. Kim, L. Wang, and A. van den Hengel, proposed that one of the primary

difficulties in learning such a metric is to ensure that the Mahalanobis matrix remains positive semi definite. Semi definite programming is sometimes used to enforce this constraint, but does not scale well [4]. J. Blitzer, K. Q. Weinberger, and L. K. Saul, proposed the metric is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a huge margin [5]. Y.Ying and P. Li proposed that the framework not only provides new insights into metric learning but also opens new avenues to the design of efficient metric learning algorithms. First-order algorithmsestablished for DMLeig and LMNN which necessity the computation of the major eigenvector of a matrix per iteration [6]. J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, proposed that this method can handle a wide variety of limitations and can optionally incorporate a prior on the distance function. Also it is fast and scalable [7].S. Chopra, R. Hadsell, and Y. LeCun, proposed that a function that maps input patterns into a target space such that the norm in the target space approximates the semantic distance in the input space. The method is applied to a face verification task [8]. A. Frome, Y. Singer, and J. Malik, proposed that a distance function for each training image as a combination of elementary distances between patch based visual features [9]. K. Q. Weinberger and L. K. Saul proposed that extended the original framework for LMNN classification in several important ways: by describing a solver that scales well to bigger data sets, by integrating metric ball trees into the training and testing procedures [10]. The system uses small amount of labeled data as training data which is used to train decision trees in hierarchy forest. These trees are used to predict the label (class) for unlabeled data. As compared to supervised learning the proposed system takes fewer amounts of labeled data for training. Also the system provides features for retrieval nearest neighbors. Nearest neighbor retrieval is done relatively faster.

III. EXISTING SYSTEM

The Mahalanobis distance is a parameter to calculate the distance between a point P and a distribution D, introduced by P. C. Mahalanobis. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. The distance is zero if P is at the mean of D, and increases as P moves away from the mean: along each principal component axis, it calculates the number of standard deviations from P to the mean of D. If every of these axes is rescaled to have unit difference, then Mahalanobis distance agrees to standard Euclidean distance in the altered space. Mahalanobis distance is unit-less and scale-invariant, and takes into account the correlations of the data set. The full hierarchy forest distance is effectively the mean of a number of weak distance functions Ht, each corresponding to one hierarchy in the forest. These distance functions, in turn, are representations of the structure of the individual hierarchies moreover the apart of two instances fall within a hierarchy, the greater the distance between them.

A. Algorithms for Existing System

Algorithms 1: HFD Learning
HFD is conceptually distinct from random forests in that the individual components of the forest represent cluster hierarchies rather than decision trees. Furthermore hierarchy forest distance also differs from the common structure of random forest in that it does not do bootstrap sampling on its training points, and its splitting functions are linear combinations rather than single-feature thresholds.

Algorithm 2: HFD Inference
Metric inference on learned HFD structures is straight forward. We feed two points to the metric and track their progress down each tree. At each node, compute associated binary linear discriminates (for root info).

Algorithm 3: Fast Approximate HFD NearestNeighbors
Comparing to a euclidean or even Mahalanobis distance. This is worsened, for many applications, by the unavailability of traditional fast approximate nearest neighbor methods, which require an explicit representation of the data in the metric space in order to function. We address the latter difficult by presenting our own fast approximate nearest-neighbor process, which takes advantage of the tree-based structure of the metric to greatly reduce the number of pairwise distance computations needed to compute a set of Nearest-neighbors for a query point x.

IV. PROPOSED SYSTEM

Here a novel tree-based non-linear metric learning method; this method can have information from both constrained and unconstrained points. Propose a relaxed constraint formulation for max-margin clustering which improves the performance of the method in hierarchical problem settings. The results show that algorithm is competitive with the state-of-the art on small- and medium-scale datasets, and superior for large-scale problems. Further a novel in-metric approximate nearest-neighbor retrieval algorithm for method that greatly decreases retrieval times for large data with little reduction in accuracy.

A. System Architecture

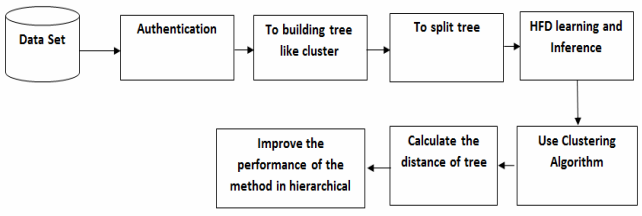

Figure: Proposed System Architecture

B. *Algorithm for Proposed System*

Module 1: HFD Learning

Module 2: HFD Inference

Module 3: Clustering Algorithm:
This algorithm will include HFD learning algorithms as well as HFD interface algorithm. The existing algorithm Fast Approximate HFD Nearest Neighbors has more time complexity so new algorithm to be developed to calculate minimum distance between neighboring points with lower time complexity.

Clustering Algorithm:

1: function Classify (X, Y, x) // X: training data, Y: class labels of X, x: unknown samples
2: for i = 1 to m do
3: Compute distance d(X, x)
4: end for
5: Compute set I containing indices for the k smallest distance d(X, x)
6: return majority label for(Y, where i∈ I)
7: end function

For determine parameter k, k is number of nearest neighbors, compute the distance among the query instance and all the training samples, sort the distance and determine, sort the distance and determine nearest neighbor based on k^{th} minimum distance, then gather the category Y of nearest neighbor. It uses the simple mainstream of the group of nearest neighbor as the guess value of the query instance.

C. *Mathematical Model for Proposed System*

$$D(a,b) = \frac{\sum Distance\ Function\ (a,b,t)}{N}$$

Distance between objects a and b is calculated as, where t is tree and N is number of trees in hierarchy forest.The objects with least distances will be returned as nearest neighbors.

V. EXPERIMENTAL SETUP AND RESULT

The proposed system manages to train the given number of trees using labeled data. These trees are used to predict the classes for unlabeled data. Each tree in hierarchy forest has its own distance function which provides distance between two nodes in the tree. The system manages to retrieve nearest neighbors as well as system proposes a novel relaxed constraint formulation for max-margin clustering which improves the performance of the method in hierarchical problem settings. The system is competitive with the state-of the-art on small- and medium-scale datasets.

Result Table:

K	Balance	Diabetes	MAGIC	Haberman
1	0.108	0.212	0.231	0.214
5	0.128	0.223	0.235	0.136
10	0.109	0.130	0.234	0.143
15	0.130	0.210	0.230	0.126

Figure: Result Table

The Following Figure shows performance of proposed system in the graph,

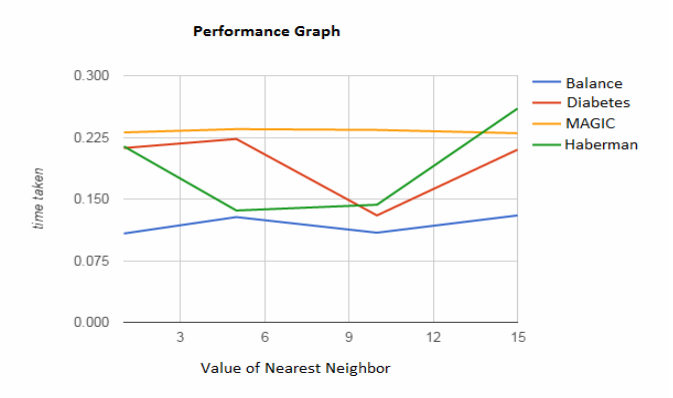

Figure . Graphical Representaion of Performance

VI. CONCLUSION AND FUTURE WORK

The proposed system a semi-supervised metric learning method based on forest of cluster hierarchies. Further propose an algorithm that can improve the performance. The algorithm shows that it can compete with currently implemented methods on small as well as large scale datasets. We have also proposed in metric approximate nearest-neighbour retrieval algorithm that reduces the retrieving time on large dataset with small compromise with accuracy.

ACKNOWLEDGMENT

I express true sense of gratitude towards my project guide Prof.V.S.Gaikwad, of computer department for his invaluable co-operation and guidance that he gave me throughout my research, for inspiring me and providing me all the lab facilities, which made this research work very convenient and easy. I would also like to express my appreciation and thanks to our HOD Prof.R.H.Kulkarni and

Director Dr.Prof.A.B.Auti and all my friends who knowingly or unknowingly have assisted me throughout my hard work.

REFERENCES

[1] D. M. Johnson, C. Xiong and J. J. Corso, “SemiSupervised Nonlinear Distance Metric Learning via Forests of Max-Margin Cluster Hierarchies,”vol. 28, no. 4, pp. 1035-1046, April 1 2016.

[2] A. Bellet, A. Habrard, and M. Sebban, “A survey on metric learning for feature vectors and structured data,”arXiv preprint arXiv:1306.6709, 2013.

[3] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, “A new kernelization framework for Mahalanobis distance learning algorithms,”Neurocomputing, vol. 73, no. 10, pp. 15701579, 2010.

[4] C. Shen, J. Kim, L. Wang, and A. van den Hengel, “Positive semidefinite metric learning with boosting,”in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 16511660.

[5] J. Blitzer, K. Q. Weinberger, and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,”in Proc. Adv. Neural Inf. Process. Syst., 2005, pp. 14731480.

[6] Y. Ying and P. Li, “Distance metric learning with eigen value optimization,”J. Mach. Learn. Res., vol. 13, pp. 126, 2012.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information theoretic metric learning,”in Proc. 24th Int. Conf. Mach. Learn., 2007, pp. 209216.

[8] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,”in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.,2005, pp. 539546.

[9] A. Frome, Y. Singer, and J. Malik, “Image retrieval and classification using local distance functions,”in Proc. Adv. Neural Inf. Process. Syst., 2006, pp. 417424.

[10] K. Q. Weinberger and L. K. Saul, “Fast solvers and efficient implementations for distance metric learning,”in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 116011.

[11] K. L. Wagstaff, S. Basu, and I. Davidson, “When is constrained clustering beneficial, and why?”Ionosphere, vol. 58, no. 60.1, pp. 6263, 2006.

AUTHORS PROFILE

Mr. Akash N. Mhetre, received the Bachelor of Engineering Degree in Computer Science & Engineering form, Dr, D Y Patil College of Engineering and technology,Kolhapur, India in year 2015. He is currently pursuing Master of Engineering from Rajshree Shahu School of Engineering and Research, JSPM NTC, Pune, India. His main research work focuses on Data Mining.

Mr. Vilas S. Gaikwad, received the BE Degree in Computer Science & Engineering from the Dr.BAMU Aurangabad, the M.Tech degree in Computer Science & Engineering from Walchand College of Engineering(An autonomous Institute), Sangli. He is currently working as Assistant Professor in the Department of Computer Engineering, Rajshree Shahu School of Engineering and Research, JSPM NTC, Pune, India. His research area includes Image Processing and Computer Network.