Open access Journal International Journal of Emerging Trends in Science and Technology

A Pairwise Homogeneity Active Learning

Authors
Sneha Mary Thomas[1], Nimmymol Manuel[2]
[1]Student, Department of Computer Science and Engineering,
Mangalam College of Engineering, Ettumanoor, Kerala, India
Email*: snehaarackal@gmail.com*
[2]Assistant Professor, Department of Computer Science and Engineering,
Mangalam College of Engineering, Ettumanoor, Kerala, India
Email*: Nimmymol.manuel@mangalam.in*

Abstract
Machine Learning has becoming an emerging topic within data mining. The active learning is also an upcoming topic. Active learning is the process of labeling unlabeled data instances by using queries and the labeling process will be done by expert labelers such as an oracle. The process of labeling will be very expensive and time consuming. The proposed method called a pairwise homogeneity active learning method which is an unsupervised label refinement method by using a pairwise homogeneity between the pair of data instances which improves the quality of the label. In this method we use a non expert labeler to provide the class label for data instances. The experimental results shows that the proposed method improves the labeling quality of the data instances which are being labeled.

Keywords: *active learning, data mining, labeling, pairwise homogeneity, data instances.*

1. Introduction

Data mining is the process of mining knowledge from large or huge amount of data. The process of machine learning is a class of data mining. In order to label a large amount of data collection it can be very difficult and highly costly process. Therefore this can be solved by using the process of active learning. The process of active learning selects the most important data instances from the huge amount of datasets and this can be used for the labeling process.

Classification is the process of finding a model such that the model defines various classes and concepts and this model can be used for prediction that is to provide the class label of data instances whose class label will be not known. The Fig.1 shows the main process of active learning. The process of classification is a two step process. The first step is a learning step and the second one is a classification process. During the learning step the training datasets are being analyzed using a classification algorithm or a classifier is build by using a classification algorithm. In the second step the test data's that is the data's which are unlabeled is being used to measure the accuracy of the classification rules. That is during this step the model is used for classification. In Fig 1 it shows the Active Learning process. The unlabeled data instances will be selected by using queries and it will be labeled by using an expert labeler such as an oracle and the labeled data instances or the set of labeled training set will be provided to the machine learning model in order to learn the model.

Labeling is the process of providing class labels to the unlabeled set of data instances. In order to provide the class label for a particular data instance the process will be time consuming therefore the proposed method which is a ulr based pairwise homogeneity method in which it uses a paiwise homogeneity information between data instances for labeling the data instances.

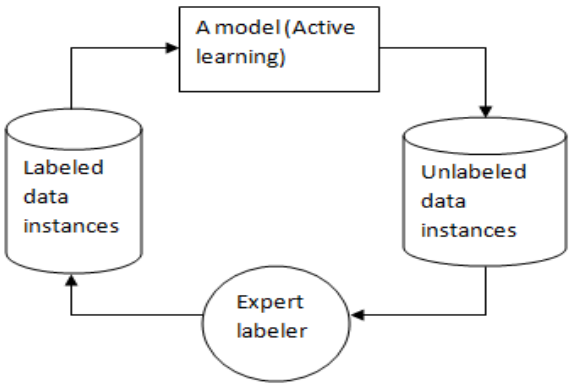

Fig 1. Process of Active Learning

The unsupervised label refinement method improves the quality of the data instances after the labeling process. By the method of active learning the most important data instances will be selected for the labeling process. During the method of active learning the expert labeler is used to provide the class label for the unlabeled data instances. But this process increases the labeling cost and the process of labeling will be a time consuming process. A Pairwise Label Homogeneity method by Yifan Fu[1] uses a non expert labeler to provide the class label to the data instances. This method reduces the labeling cost because a non expert labeler is used to provide the label to data instances. Dayong Wang[2] proposed a method which is a ULR(unsupervised label refinement) in which the quality of web facial images are refined. The main process of Active Learning is to provide the class label to the unlabeled set of data instances and labeling is the process in which the class label is being provided to the unlabeled set of data instances.

In this paper a ULR (unsupervised label refinement) based Paiwise Homogeneity method is proposed. The ULR based Pairwise Homogeneity Active Learning is used in which a pair of unlabeled set of images will be selected I norder to provide the class label of these unlabeled images using the homogeneity information[3] between these images and the unsupervised label refinement is performed in order to improve the quality of the label which is being provided to the pair of images.

The remaining sections of this paper is organized as follows. The section 2 mainly describes the related work on the Active Learning process. The section 3 describes about the proposed method and implementation. The section 4 describes experimental evaluation and section 5 provides the conclusion of this paper.

2. Related Works

Active learning is the process of selecting the most informative data instances[4] for the labeling process. The main concept behind active learning is that it uses a learning algorithm which can provide a greater accuracy which can be used for the classification purpose. The following are some of the methods that is used for the labeling process.

2.1 Pairwise Label Homogeneity

In order to manually label large amount of data collections will be very much expensive. The traditional active learning methods requires an expert labeller in order to provide the label for the data instances. Yifan Fu[1] proposed a pairwise label homogeneity based active learning method in which it uses a non expert labeller in order to provide label for the data instances. In the pairwise homogeneity method it selects a pair of data instances and non expert labeller is used to answer whether the selected pair of data instances belong to the same class or not. Here the main aim is to identify which pair of data instances should be selected for the querying. This is done by using the pairwise query on max flow path method in order to query the pair of data instances. Inorder to query the pair of data instances H. Abe[5] proposed a query learning method using bagging and boosting which uses query by bagging algorithm and query by boosting algorithm in order to select the data instances.

2.2 Optimal Data Selection

Many of the learning algorithms find out the optimal way to select the training datas.In the method that is proposed by D.Cohn[6] it mainly describes about the optimal data selection by using a feed forward neural network and two statistically based learning

architectures which are mixtures of Gaussians method and locally weighted regression method. The methods for neural network will be computationally expensive whereas the Mixtures of Gaussians method and locally weighted regression method both are accurate and efficient for optimal data selection. The mixtures of Gaussians method is a probabilistic method which assumes that all the data points will be generated from a mixture of a finite number of Gaussian distribution. The locally weighted regression is a memory based method in which the training data instances will be saved in order for future predictions. It will perform the regression using the training data instances which are mainly local to a particular point of interest.

2.3 Graph Mincut Method

Many of the learning algorithms suffer from not having enough amounts of labelled examples whereas the unlabeled examples can be collected more cheaply. A.Blum[7] proposed a method that utilizes unlabeled data for classification based on the graph cut method. This method is based on finding minimum cuts in graph and it uses the pairwise relationship between the examples. The dataset mainly consists of labelled data instances and unlabeled data instances and based on this a graph will be constructed in such a way that the minimum cut in the graph will provide the optimal binary labelling. The algorithm used in this method constructs a graph which is based on the similarity measure between the data instances and it will provide classification as output which will partition the graph in such a way it minimizes the similar pairs of data instance which will be having different labels.

2.4 Labeling Data Instance

Active learning[8] is an efficient method in order to reduce the labeling cost since the method of active learning selects the most essential data instances for the labeling process. M.Fang [9] proposed a probabilistic model which transfers knowledge from a large set of labelled instances and it makes use of a multi-labeler active learning method in which multiple labellers are used to provide class labels for the data instances. The active learning algorithm mainly selects the most informative data's and the most appropriate labeler in order to provide class labels to data instances. The process of active learning will continuously select the most important data instances.V,S Sheng[10] proposed a method which uses another label in order to improve the data quality and data mining using multiple noisy labelers.

2.5 Improve Label Quality

Dayong Wang[2] proposed a new method which is an unsupervised label refinement method which is used to improve the quality of web facial images by using the method of machine learning. It mainly uses a search based annotation method which mainly depends upon the mining weakly labeled facial images. It also used an optimization algorithm in order to make the method more feasible for real world applications. This method is an unsupervised label refinement method which is used to improve the quality of web facial images by using the method of machine learning. The web images will be having noisy and incomplete labels and in order to improve the quality of these labels the unsupervised label refinement method is being used. It has mainly used optimization algorithms for the unsupervised label refinement method.

3. Proposed Method

The proposed method mainly provides a ulr (unsupervised label refinement) based pairwise homogeneity method in which it uses an unsupervised label refinement on the data instances. The ulr based pairwise labeling method is a promising search based pairwise label scheme by mining large amount of weakly labelled facial images that will be available on the world wide web. The main working of the proposed method is provided in Fig.2. The database consists of a collection of data instances the proposed system mainly considers labeling of images. The proposed method mainly provides a ulr based pairwise label

method in which it uses a unsupervised label refinement on the data instances.

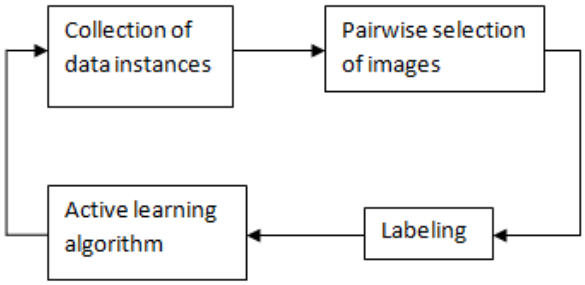

Fig 2. Working of the Proposed Method

The proposed method mainly selects most important pairs of images from the collection of data instances and based on the unsupervised label refinement method were this method may further improve the label quality [2].Based on the active learning algorithm the most important pair of images will be selected from the collection for labeling. The proposed method mainly consists of three modules: label based image search, face annotation and face annotation performance on the database. In the label based image search module, a pair of images will be selected and provided with a class label.

In Fig.2 shows the working of the Proposed Method. In the ULR(unsupervised Label Refinement) based Pairwise Homogeneity Active Learning method from the collection of unlabeled set of images a pairwise selection of images takes place and based on the similarity measure between the images the unlabeled images will be provided with labels. That is a unsupervised label refinement method is applied to thelabeled images inorder to improve the quality of the label which has been provided. The ULR(unsupervised Label Refinement) based Pairwise Homogeneity method mainly improves the quality of the labeled images.

The unsupervised learning method is the method of finding some hidden data's within an unlabeled image. The ulr which is an unsupervised label refinement method is a method in order to improve the quality of the labelled images. In ulr based pairwise active learning a pair of images or data instances will be selected and based on the similarity measure between the images or data instances the labeling procedure takes place. The proposed method mainly selects most important pairs of images from the collection of data instances and based on the unsupervised label refinement method were this method may further improve the label quality [2].

The unsupervised learning method is the method of finding some hidden data's within an unlabeled image. The ulr which is an unsupervised label refinement method is a method in order to improve the quality of the labelled images. In ulr based pairwise active learning a pair of images or data instances will be selected and based on the similarity measure between the images or data instances the labeling procedure takes place. The proposed method mainly selects most important pairs of images from the collection of data instances and based on the unsupervised label refinement method were this method may further improve the label quality [2].

4. Performance Evaluation

The system which has been developed mainly depends upon the database which consists of both labelled and unlabeled images or data instances. The fig.3 shows the graph showing the performance of the proposed system when compared with the existing system. The proposed system provides a higher quality of the labelled images when compared to the existing system.

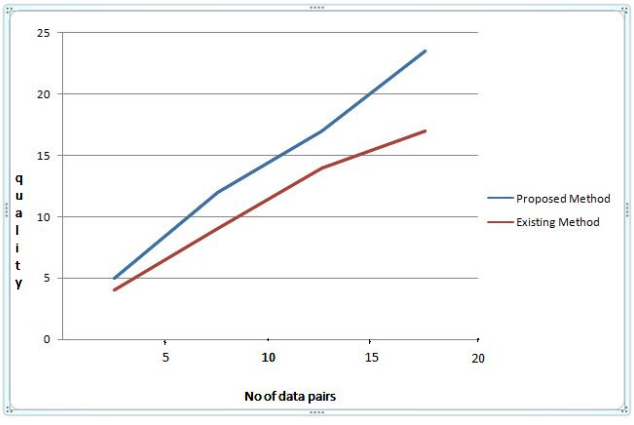

Fig 3: Comparison of Existing and Proposed Method

5. Conclusion

The proposed method mainly provides a ulr based pairwise label method in which it uses a unsupervised label refinement on the data instances. The ulr based pairwise labeling method is a promising search based pairwise label scheme by mining large amount of weakly labelled facial images that will be available on the world wide web. The performance evaluation shows that the proposed method provides much higher performance when compared to the existing systems.

References

1. Yifan Fui, Bin Li, Zingquan Zhu, "Active Learning without knowing individual instance labels:A Pairwise Label Homogeneity Query Approach",IEEE Trans.on Knowledge and Data Engineering,vol.26, no.4, April 2014
2. Dayong Wang, Steven C.H. Hoi, Ying He and Jianke Zhu, "Mining weakly labeled web facial images for search-based face annotation", IEEE Trans.on Knowledge and Data Engineering, vol.26,no.1,pp.13-64, Jan.2014
3. Y. Fu, B. Li, X. Zhu, and C. Zhang, "Do They Belong to the Same Class: Active Learning by Querying Pairwise Label Homogeneity", Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 2161-2164, 2011.
4. Y. Fu, X. Zhu, and B. Li, "A Survey on Instance Selection for Active Learning," Knowledge and Information Systems, vol. 35, pp. 249-283, 2013.
5. H. Abe and H. Mamitsuka, "Query Learning Strategies Using Boosting and Bagging", Proc. Int'l Conf. Machine Learning (ICML '98), pp. 1-9, 1998
6. D.Cohn, Z. Ghahramani and M. Jordan, "Active Learning with Statistical Models", J. Artificial Intelligence Research, vol. 4, pp. 129 145, 1996
7. A. Blum and S. Chawla, "Learning from Labeled and Unlabeled Data Using Graph Mincuts", Proc. 18th Int'l Conf. Machine Learning (ICML), pp. 19-26, 2001
8. B. Settles, "Active Learning Literature Survey," Technical Report 1648, 2009.
9. M. Fang, J. Yin, and X. Zhu, "Knowledge Transfer for Multi- Labeler Active Learning", Proc. European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Sept. 2013.
10. V.S. Sheng, F. Provost, and P.G. Ipeirotis, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2008.