THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Multiple Instances Learning Using K-Nearest Neighbor

Rina Jain
M.E., Computer Engineering, KKWIEER, Nashik, University of Pune, Maharashtra, India

Abstract:
Multi-instance learning is a variant of supervised machine learning where learner receives set of bags associated with binary label rather than each instance is labeled positive or negative. In this learning, each sample bag may has alternate feature vector that depicts it and still only one of those may be responsible for observed label of bag. So a sample is labeled as a positive bag if at least one of its instances is positive. Otherwise, it is labeled as a negative bag. MIL getting growing attention because if its suitability in numerous real world tasks such as image classification, molecular activity prediction, text or document categorization etc. In this paper, the problem definition, learning algorithm and experimental data sets related to multi-instance learning framework are briefly reviewed. The purpose of this work is to show that kNN based MIL may improved results of classification as compared other algorithms. To validate the approach Musk dataset is taken as a benchmark dataset.

Keywords: *constructive covering algorithm, kNN, Machine learning*

1. Introduction

Learning is a process that is normally associated with humans; hence the problem of designing machines that can learn falls within artificial intelligence. While invention of truly intelligent machines still seems to be long way off, many algorithms have been discovered that allow machines to make inference from observed data, effectively learning non trivial facts and behaviors. A machine learning algorithm receives examples from teacher or from the environment and attempts to learn some concept that will generalize for unobserved examples. According to ambiguity in training data, machine learning is roughly categorized into three frameworks-Supervised, Unsupervised and Reinforcement learning. Unlike to supervised learning where all training instances are with known labels, in multi-instance learning the labels of the training instances are unknown; different to unsupervised learning where all training instances are without known labels, in multi-instance learning the labels of the training bags are known and different from reinforcement learning where the labels of the training instances are delayed, in multi-instance learning there is no delay. It has been shown that learning algorithms ignoring the characteristics of multi-instance problems, such as popular decision trees and neural networks, could not work well in this scenario [1]. Traditional supervised learning deals with data which are presented to the algorithm in the form of (x, y). In multiple-instance learning (MIL), the data is provided in the form of labeled bags (B, y) where $B=\{x, ..., x_n\}$is a collection of data instances. Feature vector often referred to as instance. The relation between the label of a bag and the labels of its instances is ambiguous for the positive bags. If a bag label is positive y=1 the bag containing to the positive class. For the negative bags y=-1, there is no ambiguity and all instances belong to the negative class. So Multiple instance learning is variant of supervised machine learning where each learning examples contains a bag of instances instead of a single feature vector. Multiple instance approaches try to resolve the inherent ambiguities given in many practical learning problems. Labels are only provided for entire bags and the task is to learn a model that predicts the classification labels for unseen future bags**.** Machines are far more efficient and reliable than humans at processing large amounts of data. MIL has received considerable amount of attention due to both its theoretical interest and type of representation fits for a number of real-world learning scenarios e.g. drug activity prediction[1]**,**text categorization[7], image classification[18],object detection in images[14],content based image classification[15],visual tracking [16], computer security[17] ,web mining[10],spam filtering[9] etc. Rest of paper is organized as follows. Section 2 presents survey of literature along with pros and cons of some the existing methods. MIL algorithm with the main contribution of this paper is described in section 3. Section 4 reports the data sets and results. Finally, section 5 summarizes this paper and raises issues for future work.

2. Existing Work

To solve the multiple instances problem large number of algorithms is proposed. Moreover they are also classified into bag level and instance level algorithms. Bag level methods focus on bags label rather than knowing the labels of instances in it, whereas other method works on instances and then combine the labels of instances to obtain label of unknown bag. The term multi-instance learning was introduced by Dietterich et al. [1] when they were investigating the problem of drug activity prediction.They proposed three axis-parallel rectangle (APR) algorithms to solve the drug activity prediction problem, which tries to search for appropriate axis-parallel rectangles constructed by the conjunction of the features They create an asymmetric assumption

regarding the process that determines class labels of bag based on instances in the bag. Many algorithms use that as standard assumption. DD algorithm is proposed by Maron et al. [2] suggested a concept point, that depicts a portion of instance space that is dense w.r.t. instances from positive bags. A few years later, DD algorithm extended further by adding it with the EM (Expectation-Maximization) algorithm, resulting in EM-DD algorithm proposed by Zhang et al. [5]. Natural scene classification, stock selection, drug activity prediction, image retrieval are some application of DD and its extension EMDD. In 2002, to slove MIL problems, Andrews et al. [7] suggests two methods to exploit the standard Support Vector Machine. The purpose was to point out the maximum-margin multiple-instance separating hyper plane in which at least one positive instance from all positive bags was located on the other side of hyper plane and all instances in each negative bag were located on other side. In 2006, Zhang et al. [8] proposed RBF-MIP algorithm, which is derived from the well-known Radial Basis Function (RBF). Wang et al. [4] proposed a lazy learning approach using kNN algorithm that in turn uses Hausdorff distance for measuring the distance between set of point. Two variants of this method, Bayesian KNN and Citation KNN were proposed in [2]. Deselaers and Ferrari [13] uses conditional random field where bag treated as nodes and instances treated as states of node. Babenko et al. [12] proposed bag as manifolds in the instance space. Recently, Jiang et al. [19] suggested improved version of lazy learning kNN algorithm as Bayesian Citation-kNN (BCkNN) algorithm. All of the aforesaid methods were developed using either the probabilistic EM, SVM or nearest neighbor approach. Although several algorithm efficiently worked, it is noteworthy that closely all these algorithms referenced above try to choose the positive bags or true positive instances in the positive bag instead of eliminating the false positive instances. Because of that, these methods become very complex and often take long time to run. In order to enhance the classification accuracy and reduce the complexity of algorithm, proposed system suggests the constructive covering algorithm with kNN algorithm to generate a set of covers to eliminate the false positive instances.

3. Multi- Instance Learning

Multi-instance learning, as defined by Dietterich et al. [1], is a variation on the standard supervised machine learning scenario. In MI learning, each example consists of a multiset (bag) of instances. Each bag has a class label, but the instances themselves are not explicitly labeled. The learning problem is to build a model based on given example bags that can accurately predict the class labels of future bags. An example will guide to illuminate the concept. Chevaleyre & Zucker [21] refer to this example as the simple jailer problem. Imagine that there is a locked door, and we have N keychains, each containing a bunch of keys. If a keychain (i.e. bag) contains a key (i.e. instance) that can unlock the door, that keychain is considered to be useful. The learning problem is to build a model that can predict whether a given keychain is useful or not.

3.1. Definition

Let X is input space and Y= {0, 1} be the class label, binary output space. F:X→Y is a learning function for traditional supervised learning, form a set of training instances $\{(x_1,y_1)(x_2,y_2),..,(x_m,y_m)\}$Where $x_i \in X$ is one instance and $y_i \in Y$ is label associated with x_i. In MIL, $\{(B_1,y_1)(B_2,y_2),..(B_m,y_m)\}$ is a training set of m labeled bags.Given a dataset D, instances in bag B_i defined as $\{x_1,x_2,,..x_n\}$. Let d be dimension of x. Now, D^+ and D^- denotes all instances of positive and negative bags resp. where $D^+ =\{x_i^+ |i=1,2,..,p\}$, $D- =\{x_j^+ |j=1,2,..,n\}$ and $D=D^+U\ D^-$. In this each instance belongs to one specific bag. So, $B^i\ U\ B^j=\phi$.

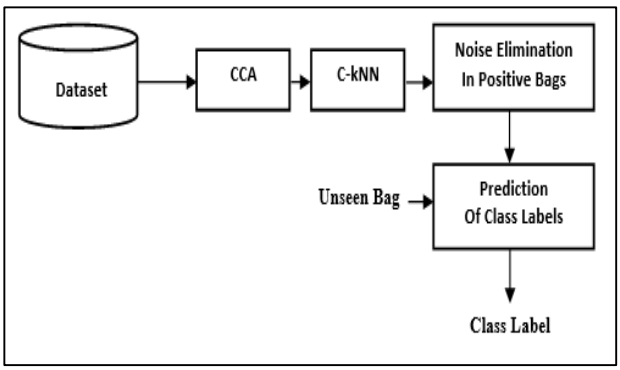

Figure 1: Architecture of Multi-instance classification learning

3.1. K-Nearest Neighbor Algorithm

kNN is widely used learning algorithm and well known for its relatively simple implementation and decent results. The main idea of kNN algorithm is to find a set of k objects in the training data that are close to the test pattern, and base the assignment of a label on the predominance of a particular class in this neighbor. kNN is a lazy learning technique based on voting and distances of the k nearest neighbors. Given training set D and a test pattern x, kNN computes the similarity (distance) between x and the nearest k neighbors. The label of x is assigned by voting from the majority of neighbors. In proposed work three different kNN algorithms used to compare accuracy and computation time between them. They are - Bayesian-kNN, Citation-kNN, and Bayesian-Citation kNN (BCkNN)[19]. Bayesian-kNN uses probabilistic approach to calculate its neighbor while Citation-kNN algorithm not only takes into account the neighbors (references) of bag b but also the bags that count b as a neighbor (citers). Whereas BCkNN firstly finds its r references and c citers, then a Bayesian approach is applied to its r references and a distance weighted majority vote approach is applied to its c citers.

3.2. Constructive Covering Algorithm

Zhang and Zhang [6] proposed supervised learning algorithm for McCulloch-Pitts neural model named as Constructive Cover Algorithm CCA). To reorganize the structure of bags Zhao et. al [12] uses the cover set as the new structure of bag. So that the false positive instances can be excluded. The main idea of CCA is map excluded. The main idea of CCA is mapping all instances in the data set to a d-dimensional sphere S^d at first. The sphere with an instance as center and r as radius is represent as covers. Their sphere neighbors (covers) are utilized to divide the patterns (instances).

First data transformed using T(x) = (x, $\sqrt{R^2 - \|x\|^2}$, R≥max {||x|| | x ϵ D}) such that x is random instance and R is the greater or equal to maximum value of all instances. Transformation T: D→S^d, where S^d is a d-dimensional sphere of the d+1 dimensional space, $\sqrt{R^2 \quad \|x\|^2}$is the additional value of x.

After that, series of positive covers that only consists of instances in the positive bags and series of negative covers that only consists of instances of negative bags are constructed. To generate covers, first of all, an instance x_i ϵ D selected randomly. Consider, X be the set of instances has the same label as x_i and -X the set of instances having opposite label from x_i. Then distance d_1 and d_2 computed such that

d_1= max {<x_i, x_j> | x_i ϵ X, x_j ϵ -X},

d_2= min {<x_i, x_k> | x_i, x_k ϵ X}

Here x_j is the closest instance from x_i which belongs to set off-X, whereas x_k is furthest instance from x_i which belongs to set of X. d_2 must be smaller than d_1 and where <x_1,x_2> denote the inner product between instances x_1 and x_2. Note that smaller the distance bigger the inner product. Next, radius r of sphere neighbor is calculated as r = (d_1+d_2)/2. The result of CCA is a series of covers, each of which contain samples belonging to the same class.

3.3. Noise Elimination in Positive Bags

For eliminating false positive instances kNN algorithm is utilized on cover obtained by CCA. For each $PCover_i$, its nearest neighbor calculated and checks if majority of its neighbors are belongs to set NCover, then it added in Ncover and deleted from NCover. The distance between two cover covers is calculated using Hausdroff distance (HD).

MI data set transformed into positive cover set (PCover) and negative cover set (NCover). Fair amount of noises in positive bags are excluded.

3.4. Predicting Class Labels of Test Bags

In this method, a $PCover_i$, and $NCover_j$ is treated as the new structure of bag. Large numbers of noises in the positive bags are excluded during above procedures, it's now quite convenient to predict the labels of test bags using kNN algorithm at bag-level. It estimates the resemblance between each test bag and its nearest neighbors, if there are more negative covers around a test bag, then bag labeled as negative otherwise positive.

4. Results and Discussion

Multiple-instance classification learning algorithm eliminates the false positive instances at cover-level and labels the unknown bags at the bag-level. Two real-world benchmark data sets – Musk data sets (http://archive.ics.uci.edu/ml/datasets/Musk+%28Version+2%29) i.e. Musk1 and Musk 2 are used for experiments. Dataset contains different feature vectors of molecules and their class label. In this case, if molecule binds to target protein (putative receptor in human nose), then it smells like a musk. For determining whether molecule and target protein bind, shape of molecule is an important factor. However molecules are flexible and exhibit a wide range of shapes. Each molecule is represented by a bag and the bag's label is positive i.e. musky if molecule binds well to target protein. A bag made up of instances, where each instance represents one formation i.e. shape that molecule can take. After learning, it returns a concept which tells constraints on the shape of molecule that would bind to the target protein.

Data set	Total bags	Number of Positive bags	Number of Negative bags	Avg. instances per bag
Musk 1	92	47	45	5.17
Musk 2	102	39	63	64.69

Table 1: Summary of the Two Musk Datasets

Characteristics of datasets described in table 1.Each conformation represented by feature i.e. ray representation described in [1]. Musk 2 contains molecule that have more possible conformations i.e. instances than Musk1 is the main difference between two datasets.

.

5. Conclusion

The multi-instance problem is a variant of supervised problem, arises in real world tasks where the examples are ambiguous, single object may have many alternative feature vectors that represent it and yet only one of those feature vectors may be responsible for the observed classification of object. CCA is used to break through and restructure the original bags into covers so that noises in the bags can be excluded by using various kNN algorithms. Then, covers as a whole, determines the labels of the unknown bags. So this is a cover- level kNN algorithm, differ from previous bag or instance-level algorithm.

Discovery of different domains, where this algorithm can be applicable and suitable is one direction of future work. In addition, whether MIL problem can be converted into a supervised problem is another direction of future work.

6. References

1. T. G. Dietterich, R. H. Lathrop, and T. Lozano-P erez, solve the multiple Instance problems with axis-parallel rectangles, Artificial Intelligence, vol. 89, no. 1, p31-71, 1997.
2. O. Maron and T. Lozano-P erez, a framework for multiple instance learning, Advances in Neural Information Processing Systems, vol. 10, pp. 570-576, 1998.
3. O. Maron, Learning from ambiguity, Ph.D. dissertation, Massachusetts Institute Technology, USA, 1998.
4. L. Zhang and B. Zhang, A geometrical representation of mcculloch-pitts neural model and its applications, IEEE Transactions on Neural Networks , vol. 10, no. 4, pp. 925- 929, 1999
5. J. Wang and J. D. Zucker, Solving the multiple-instance problem: A lazy learning approach, in Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 1119-1125, 2000.
6. Q. Zhang and S. A. Goldman, EM-DD: An improved multiple-instance learning technique, Advances in Neural Information Processing Systems, vol. 14, no. 2022, p1073-1080, 2001.
7. S. Andrews, I. Tsochantaridis, and T. Hofmann, Support vector machines for multiple-instance learning, Advances in Neural Information Processing Systems, vol. 15, pp. 561-568, 2002.
8. M. L. Zhang and Z. H. Zhou, Adapting RBF neural networks to multi-instance learning, Neural Processing Letters, vol. 23, no. 1, pp. 1-26, 2006.
9. Z. Jorgensen, Y. Zhou, and M. Inge, A multiple instance learning strategy for combating good word attacks on spam filters, The Journal of Machine Learning Research, vol. 9, no. 6, pp. 1115-1146, 2008.
10. B. B. Ni, Z. Song, and S. C. Yan, Web image mining towards universal age estimator, in Proceedings of the 17th ACM International Conference on MultimediaInt, ACM,pp. 85-94, 2009.
11. T. Deselaers and V. Ferrari, A conditional random field for multiple-instance learning, in Proceedings of the 27th International Conference on Machine Learning, Morga Kaufmann Publishers Inc,pp. 1119-1125,2010.
12. B. Babenko, N. Verma, P. Dollar, and S. J. Belongie, Multiple instance learning with manifold bags, in Proceedings of the 28th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 81-88, 2011.
13. D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence, Multiple instance learning on structured data, Advances in Neural Information Processing Systems, vol. 24, pp.145-153, 2011.
14. Z. Q. Qi, Y. T. Xu, L. S. Wang, and Y. Song,Online multiple instance boosting for object detection, Neurocomputing, vol. 74, no. 10, pp. 1769-1775, 2011.
15. Z. Y. Fu, A. Robles-Kelly, and J. Zhou, MILIS:Multiple instance learning with instance selection,IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 958-977, 2011.
16. Y. Xie, Y. Y. Qu, C. H. Li, and W. S Zhang, Online multiple instance gradient feature selection for robust visual tracking, Pattern Recognition Letters, vol. 33, no. 9, pp. 1075-1082, 2012.
17. M. Bellare, T. Ristenpart, and S. Tessaro, Multi-instance security and its application to password-based Cryptography, Advances in Cryptology-CRYPTO 2012, Springer, pp. 312-329, 2012.
18. D. T. Nguyen, C. D. Nguyen, R. Hargraves, L. A. Kurgan,and K. J. Cios, mi-DS: Multiple-instance learning algorithm, IEEE Transactions on Systems, Man, and Cybernetics Society. Part B, Cybernetics, vol. 43, no. 1,pp. 143-154, Feb. 2013.
19. L. X. Jiang, Z. H. Cai, D. H. Wang, and H. Zhang, Bayesian citation-KNN with dis-tance weighting, International Journal of Machine Learning and Cybernetics, pp. 1-7,2013.
20. Shu Zhao, Chen Rui, and Yanping Zhang, MICkNN : Multi-Instance Covering kNN Algorithm, TSINGHUA SCIENCE AND TECHNOLOGY,vo;. 18,no. 4, pp.360-368,2013.
21. Chevaleyre, Y. & Zucker, J.-D. 2001. Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. Application to the mutagenesis problem. Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies Of Intelligence, Springer, pp.204-214