Study of Different Multi-instance Learning kNN Algorithms

Rina S. Jain,
Department of Computer Engineering,
K.K.W.I.E.E.R., Nashik,
University of Pune, India

Abstract: Because of it is applicability in various field, multi-instance learning or multi-instance problem becoming more popular in machine learning research field. Different from supervised learning, multi-instance learning related to the problem of classifying an unknown bag into positive or negative label such that labels of instances of bags are ambiguous. This paper uses and study three different k-nearest neighbor algorithm namely Bayesian -kNN, citation -kNN and Bayesian Citation -kNN algorithm for solving multi-instance problem. Similarity between two bags is measured using Hausdroff distance. To overcome the problem of false positive instances constructive covering algorithm used. Also the problem definition, learning algorithm and experimental data sets related to multi-instance learning framework are briefly reviewed in this paper.

Keywords: Bayesian kNN, citation kNN, constructive covering algorithm, Machine learning, Multi-instance problem

1. INTRODUCTION

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. Figure 1 shows ambiguity spectrum for learning framework. The prediction tasks for MI problems are more difficult than those for single-instance learning problems because the real cause of the bag label is ambiguous. 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].

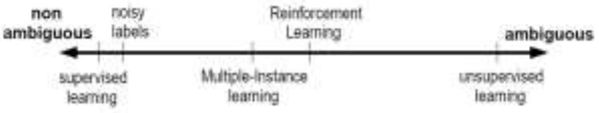
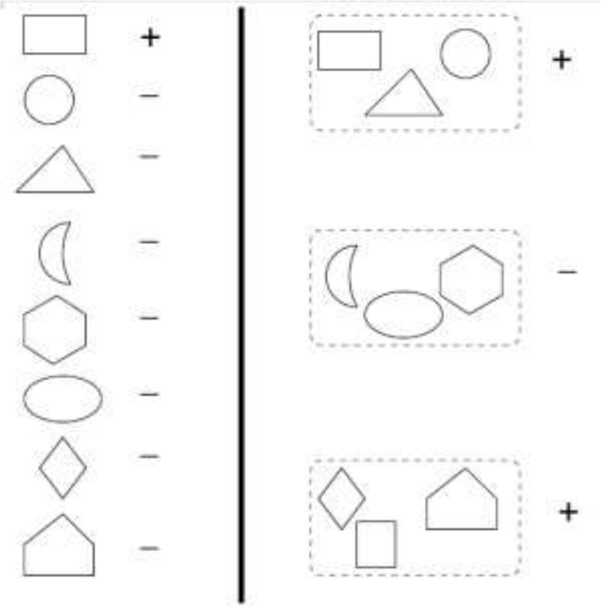


Figure. 1 Ambiguity spectrum (by O. Maron et al. [3])

In multiple-instance learning, the input consists of labeled examples (called bags) consisting of multisets of instances, each described by an attribute vector, training set comprises of such bags and task is to predict the labels of unobserved bags. Figure 2 (by S Ray et al.) illustrates the relationships between standard supervised learning and multiple-instance learning to get clear idea about them. Consider the example "whether figure is rectangle or contain at least rectangle", Label can be positive (+) or negative (-), as for two-class classification problem. (a) In supervised learning, each example (geometric figure) is labeled. A possible concept that explains the example labels shown is "the figure is a rectangle". (b) In MI learning, bags of examples are labeled. A possible concept that explains the bag labels shown is "the bag contains at least one figure that is a rectangle."



a) Supervised Learning b) Multi-Instance Learning

Figure. 2 Example depicting Relationship between supervised and multi- instance learning

So, in MIL, if bag comprise of at least one positive instance which represent the output then bag is labeled as positive. If bag consists of all negative instances then it labeled as negative.

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. RELATED WORK

The multi-instance concept was first formally introduced by Dietterich et al. [1]. It was originally inspired by a drug activity prediction problem. In this task, a molecule can have several conformations (i.e. shapes) with different properties that result in the molecule being of "musk" or "non-musk" type. However, it is unidentified which particular conformation is the cause of a molecule being of the "musk" type. The conventional single-instance setting (Supervised learning) cannot represent this application problem properly as one molecule may have several other conformations. Therefore, Dietterich et al. [1] proposed the multi-instance setting, in which each sample is represented by a collection of single instances instead of a single instance and also made an asymmetric assumption regarding the process that decides class labels of bag based on instances in the bag. Many algorithms use that as standard assumption. DD algorithm is projected by Maron et al. [2] that includes a concept point that describes a portion of instance space that is dense w.r.t. instances from positive bags. A few years later, DD algorithm stretched further by adding it with the EM (Expectation-Maximization) algorithm, resulting in EM-DD algorithm proposed by Zhang et al. [5]. DD and its extension EMDD have been used on various MIL problems such as stock selection, drug activity prediction, natural scene classification and image retrieval.

In 2002, to solve MIL problems, Andrews et al. [7] advises two methods to exploit the standard Support Vector Machine. The aim was to point out the maximum-margin multiple-instance separating hyper plane in which at least one positive instance from all positive bags was placed 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] recommended RBF-MIP algorithm, which is derived from the well-known Radial Basis Function (RBF).

Wang et al. [4] suggested 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.

The experimental results also show that different MI algorithms are appropriate for different MI problems and that no single MI algorithm was well-suited to every MI domain that was tested. In order to improve the classification accuracy and lessen the complexity of algorithm, proposed system suggests and compare the constructive covering algorithm with different kNN algorithm to create a set of covers to exclude the false positive instances.

3. MULTI- INSTANCE LEARNING

Multi-instance learning, as well-defined by Dietterich et al. [1], is a variation on the standard supervised machine learning scenario. In MI learning, every example consists of a multiset (bag) of instances. Each bag has a class label, but the instances themselves are not directly labeled. The learning problem is to form a model based on given example bags that can precisely predict the class labels of future bags. An example will help to illuminate the concept. Chevaleyre et al. [21] refer to this example as the simple jailer problem. Visualize that there is a locked door, and has 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 of MIL

Let X be input space and Y= {0, 1} be the class label, binary output space. $F:2^X \rightarrow Y$ 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 Bi 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, $B_i U B_j=\phi$.

Block diagram of Multi-instance learning is explained below.

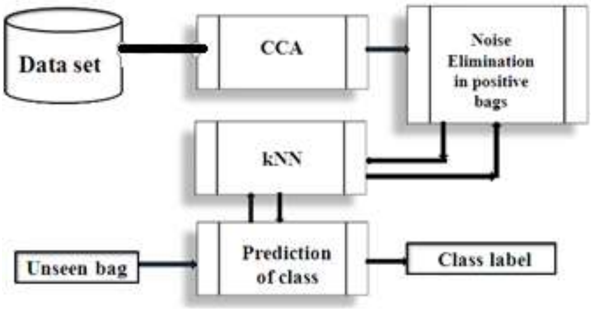


Figure. 2 Block diagram of MIL system

3.2 Constructive Covering Algorithm (CCA)

Supervised learning algorithm for McCulloch-Pitts neural model named as Constructive Cover Algorithm (CCA) proposed by Zhang and Zhang [6].This algorithm act as backbone of the system as it reorganize the structure of bags and uses the cover set as the new structure of bag. This algorithm is transformed by Zhao et al. [16] so that it can be useful in multi-instance concept. This algorithm finds and eliminate the false positive instances. The main idea of CCA is mapping all instances in the data set to a d-dimensional sphere S^d at first. Cover set is the final output and cover is nothing but a sphere with an instance as center and r as radius.

First data converted using $T(x) = (x, \sqrt{R^2 - ||x||^2}$, $R \geq \max \{||x|| \mid x \in D\}$) such that x is random instance and R is the greater or equal to maximum value of all instances. Transformation $T: D \rightarrow S^d$, where S^d is a d-dimensional sphere of the d+1 dimensional space, $\sqrt{R^2 - ||x||^2}$ is the additional value of x. After that, sequence of positive covers that only consists of instances from the positive bags and sequence of negative covers that only consists of instances of negative bags are constructed. To generate covers, first of all, an instance $x_i \in D$ selected arbitrarily. 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

$d_1= \max \{<x_i, x_j> \mid x_i \in X, x_j \in -X \}$,

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 of X, whereas x_k is furthest instance from x_i which belongs to set of X. d2 must be smaller than d1 and where <x_1,x_2> signify 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 K Nearest Neighbor Algorithm (kNN)

kNN is widely used learning algorithm and well known for its relatively simple execution 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. But rather than Euclidean distance, Hausdorff distance is used to measure similarity between two covers. Wang et al. [5] presented two types of Hausdorff distance that are - maximal Hausdorff distance (maxHD) and minimal Hausdorff distance (minHD).Given two sets of instances X={x_1,x_2,…x_n} and Z={z_1,z_2,…,z_n}, the maxHD defined as -

maxHD(X,Z)=max{h(X,Z),h(Z,X)}

where $h(X,Z)=\max_{x \epsilon X} \min_{z \epsilon Z} || x-z||$

The minHD is defined as -

minHD(X,Z)=min|| x-z ||

where || x-z || is the Euclidean distance between instance x and y.

Proposed work uses and compares accuracy and computation time of three kNN algorithm of this system. They are-

I. Citation kNN[5]:- C-kNN algorithm not only takes into account the k neighbors (references) of bag b but also the bags that count b as a neighbor (citers).Where number of citers c set to k+2 ,same as in [5] .

II. Bayesian kNN [5]:- Bayesian approach provides probabilistic approach that calculates explicit probability of hypotheses. For each hypothesis y that the class of b can take, the posterior probability of y is p(y| {y_1,y_2,..y_k }).According to the Bayes theorem, the maximally probable hypothesis is:

$$\arg\max_{c} p(y|\ \{y_1,y_2,..y_k\})=\arg\max_{c} p(\{y_1,y_2,..y_k\}|y)p(c).$$

where y_i is either positive or negative,

III Bayesian-Citation-kNN (BCKNN) [19]:- It is combined approach of Bayesian and distance weighting where firstly, Bayesian approach is applied to its k references and then distance weighting approach is applied to its c citers.

3.4 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 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.

3.5 Predication of Class label 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.

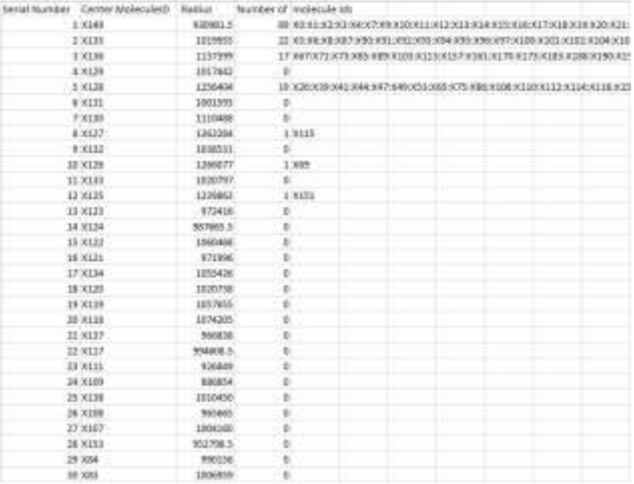
4. RESULTS AND DISCUSSION

The 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.

TABEL I
SUMMARY OF THE TWO MUSK DATA SETS

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

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



Serial Number	Center MoleculeID	Radius	Number of	molecule ids
1	X149	920981.5	[illegible]	X0:X1:X2:X3:X4:X7:X9:X10:X11:X12:X13:X14:X15:X16:X17:X18:X19:X20:X21:
2	X135	1019955	22	[illegible]
3	X136	1157399	17	[illegible]
4	X126	1017443	0	
5	X128	1256404	19	[illegible]
6	X131	1001593	0	
7	X130	1110488	0	
8	X127	1262284	1	X115
9	X132	1038511	0	
10	X129	1266077	1	X85
11	X143	1020797	0	
12	X125	1226862	1	X151
13	X123	972416	0	
14	X124	987865.5	0	
15	X122	1060468	0	
16	X121	971996	0	
17	X134	1055426	0	
18	X120	1020758	0	
19	X119	1057655	0	
20	X118	1074205	0	
21	X137	986838	0	
22	X117	994808.5	0	
23	X111	926849	0	
24	X109	886854	0	
25	X138	1010450	0	
26	X108	993465	0	
27	X107	1004160	0	
28	X153	952708.3	0	
29	X84	990156	0	
30	X83	1006939	0	

Figure 3 Snapshot of output file of CCA

Figure 3 shows output file of CCA which depicts some cover contains many instances while some none.

5. CONCLUSION

The multi-instance problem is the extension of supervised problem, arises in real world tasks where the samples are ambiguous, single example may has 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 multi-instance kNN learning algorithm, differ from previous bag or instance-level algorithm.

Detection of different fields, where this algorithm can be appropriate and suitable is one direction of future work. In addition, whether Multi-instance learning problem can be transformed into a supervised problem is additional direction of future work.

6. REFERENCES

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-P erez, Solving the multiple Instance problem 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 17^{th} 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 distance 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.