电子科技大学

硕士文献阅读综述

班 学 号：201521060324

姓 名：黄晨

指导教师：邵俊明 教授

学科专业：计算机软件与理论

所在学院：计算机科学与工程学院

指导教师阅读后签字：

摘 要

在信息爆炸的今天，人们渐渐淹没在形形色色的数据海洋之中，如火如荼的数据挖掘和机器学习，发挥着越来越重要的作用，尤其兴盛在计算机视觉，自然语言处理和深度学习领域。然而，在数据挖掘技术在这些领域成功应用的背后需要所有数据都被完全标注，即标签数据，这意味着，为了从杂乱的数据中发现更多有价值的信息，我们需要对所有的数据进行人工标注，对数据进行监督学习。然而，现实中的巨大数据往往是没有被标注的，人工标注数据的工作将是极度耗费人力物力的，并且现代生活中的数据产生方式已经发生了巨大的变化，各个领域的数据往往以流式数据的形式呈现，使得人工数据标注将变成一项“更加长期”的工作，这往往是不现实的。但事实上，由于半监督学习的提出，允许人们只用对其中的很少一部分数据进行标注。半监督学习旨在应对只含有少部分标记数据的挖掘任务，即同时利用大量无标签数据和少量有标签数据，建立一个具有更好泛化能力的模型，使得其比监督学习具有更好的效果，因此被广泛应用于分类，聚类，回归分析和数据流挖掘中。半监督学习在现实中的应用已经取得很多成功，但是由于现实世界的数据复杂多变，半监督学习同样也存在问题，即使用大量无标签数据之后建立的模型效果，比只使用标签数据之后建立的模型效果还差，这一现象可以被称为半监督学习的可靠性问题[40][45]。因此，建立一种以更加安全的方式来使用无标签数据的模型是一项重要的科学任务，同样也有很强的应用价值。

本报告研究首先介绍了当前半监督学习常用的几类模型，包括协同训练模型，最大间隔模型, 基于图的模型, 在线的半监督模型以及其他模型；然后分析了当前可靠的半监督学习的几类主要方法，包括基于集成框架的模型，基于半监督与监督方法的权衡模型和基于自适应权重的方法；最后本报告详细的总结分析了现有可靠性半监督学习模型的不足之处。

关键词：半监督学习，可靠性，数据流

1、国内外研究现状

早在 20 世纪 60 年代开始，提出的自训练模型[1]便开启了半监督学习的历史，渐渐的受到了国内国际众多研究机构的极大关注。截止今天许许多多的半监督学习模型被相继提出，并且在时间序列分析[48][49]，深度学习[50][53]，医疗诊断[59]等等应用中也得到广泛的应用

1.1 半监督学习

在数据挖掘和机器学习中，模型建立的背后往往有着不同的假设。而半监督学习基本假设在于：流型假设(Manifold Assumption)和聚类假设(Cluster Assumption)。前者假设认为高维数据存在与一个低维流型上，在同一个流型上的数据更有可能具有相同的标签；而后者假设认为相似的数据在同一个簇中，这些数据倾向于拥有相似的标签。通常来说，半监督学习，按照学习的目标，可以分为两类：感应式学习(Inductive)和直推式学习(Transductive)。后者只关注模型在训练集上的效果，即关注训练数据集中的无标签数据的预测问题，而前者还关注于未来的测试数据，即考虑在整个 X 的空间上有一个很好的泛化能力。

现有的半监督学习模型可以大致分为：协同训练模型和多视角模型 (Co-training model & Multi-view model)，最大间隔模型(Max-margin model), 基于图的模型(Graph-based), 在线的半监督模型(Online SSL model)以及其他模型。

（1） 基于图的模型

基于图的半监督模型是一个现今最常用的模型，它往往是直推式的，并且是基于流型假设的，它的主要任务在于图的构建和在图上的标签预测问题。其中，图的构造问题对该类模型起着比较关键的作用，它直接影响了模型的标签传播效果。其构造方式往往是无向图，可以基于全连接图，KNN 图，a-邻域，b-matching[2]和 L1-graph[3]，又有文献提出采用有向图[4]的方式。而对图上的标签预测建模主要从两个角度出发，一个是对有标签数据的拟合，另一个是对所有数据上的光滑性惩罚。前者往往考虑预测模型在有标签数据上的平方损失，而后者通常用图的拉普拉斯作为光滑性正则项进行建模，如基于组合图的拉普拉斯[5][6]，归一化拉普拉斯图[7]，指数拉普拉斯图[8]和树形拉普拉斯[9]，另外也有文献采用其他正则项来惩罚模型的光滑性：局部线性正则[10][55][56]，局部学习正则项[11]和局部与全局统一的正则项[12]。此外，文献[54]在流型假设的基础上，通过衡量数据的局部协方差矩阵之间的 Hellinger 距离，拟解决多个流型重叠的情况下的半监督分类问题。

（2） 最大间隔模型

传统的最大间隔模型（如 SVM）不同，其单纯的旨在寻找能将训练数据集一分为二的最适超平面，而 S3VM[13][51]和 TSVM[14][52]等等基于最大间隔的半监督模型的不同之处在于，最大化分类间隔同时，还需要考虑无标签数据和有标签数据，获取整个数据集上的最小分类损失和最大化标签光滑程度；文献[15]在 TSVM 的背景下，从“正则力度”的角度，分析了无标签数据的作用，随着力度的由小到大，模型渐渐从监督模型变化到半监督模型；文献[30]基于最大间隔假设从未标记数据中有效提取信息，以估计贝叶斯决策边界进行分类。

（3） 协同训练模型和多视角模型

协同训练模型是半监督学习中较为常用的算法[16][17][18][69]，它的思想在于，用现有训练模型评估训练集中无标签数据，并反馈地修正该训练模型。首先使用标签数据初始化模型 f，然后使

用 f 对无标签数据进行分类，评估分类结果，选择少部分置信度高的无标签数据和他们的预测标签，加入训练数据集中，进行下一次循环。该算法不断修正训练模型，直到满足终止条件，使得各个分类器在训练数据上达到最大程度的统一。而文献[60][61][62][63]使用了多视角模型，旨在使得不同视角建立的模型在训练数据上的预测达到一致。由于这一大类的模型，具有“错上加错”的缺点，模型的勿分类会在下一次迭代中进一步扩大，所以协同训练模型和多视角模型渐渐淡出了人们的视线。

（4） 在线的半监督模型

在线的半监督模型旨在顺应时代背景，解决数据流挖掘问题，文献[21]基于流型假设，对一定时间窗口内的数据建立模型，通过使用缓冲技术和在线随机投影树的技巧实现模型随着数据流而动态更新；文献[22]维护数据流上的多个动态微簇结构，并且通过衡量微簇的标签纯度来更新微簇集合，然后采用基于微簇的权重投票的机制进行标签预测。

（5） 其他模型

文献[23]提出了一个通用于任何监督分类器的 boosting 框架，文献[24][25][26]采用和 LDA 的思路做半监督分类，用有标签数据来获得判别信息，用所有数据来获取固有的数据结构信息，文献[27][28]从非负矩阵分解的方法入手处理半监督分类问题；文献[29]基于电场理论，把标签数据看做是点电荷，无标签数据的标签等于那个点处的电势能，并结合标签传播的思想建立模型。文献[19]是基于图上的生成模型，它旨在将有标签数据和无标签数据通过共享分布参数联系起来，将类的条件概率通过标签传播来估计，类的先验通过线性回归来估计。文献[57]使用无标签数据的稀疏表征来构造标签数据的稀疏表征，然后在新标签数据的表征下构建 SVM 模型作为最终的半监督模型。

1.2 可靠的半监督学习

需要注意的是，加入无标签后并不是都有帮助的，有时候甚至会降低学习效果。产生这个现象的一个简单的解释是，虽然半监督学习假设无标签数据和标签数据都服从于同一种采样分布，而无标签数据的引入旨在补足少量的标签数据，但训练数据中仍然会存在采样偏差，使得无标签数据和标签数据的训练数据分布存在差异，而更困难的情况在于，标签数据和无标签数据的采样分布不一样，而存在固有的不同，无标签数据的分布$\mathrm{P}(X_u)$并不一定是$\mathrm{P}(X_u, X_l, \mathrm{Y})$的边缘分布；另一方面，半监督学习对数据做出了很多假设，如果实际上数据并不满足这些假设，模型效果自然会受到损害，这种种的情况使得在实际使用半监督模型时，我们无法得知无标签数据是否真实扩充了仅有的标签数据，不加考虑得直接引入无标签数据可能会导致模型性能的下降[40][45][65]，文献[66][67][68]进一步从理论上分析了，引入无标签数据之后的模型分类误差的界。因此，安全的半监督学习算法受到了越来越多的关注，旨在达到引入无标签数据之后的模型效果至少会不低于原模型效果的目的。

为了达到引入无标签数据之后的模型效果至少会不低于原模型效果的目的，安全可靠的半监督学习方法渐渐引起人们的注意。当前学者们的方法主要可以分为三类：基于集成框架的模型，基于半监督与监督方法的权衡模型和基于自适应权重的方法。另外也有文献从其他角度出发，比如文献[41][42]通过放松协同训练模型的条件独立性假设，以达到可靠分类的效果。文献[46]以平方损失建立模型，通过在监督模型的解的基础上，对解进行投影，找到一个比监督模型解，更加靠近真实解的一个半监督模型的解，并且保证这个解的平方损失一定比原监督模型小。同时文献[58]提出了一个半监督线性回归模型，证明在 X 的一阶矩和二阶矩已知的情况下，该模型的解以一般的线性回归模型有更好的

效果。

（1） 基于集成框架的模型

基于集成框架的模型的特点在于使用多个初始的半监督模型和一个监督模型。首先对这些不同模型在数据集上的预测效果进行综合分析，得出最终的预测结果。文献[31]通过对同一无标签数据点在不同模型输出的不一致性进行建模，分析该点的风险度，进而将置信度不高的无标签数据点由监督分类器进行预测；文献[32]假设一个可靠的预测结果，应该是在不同模型输出上具有最大间隔，而间隔内的不置信数据结果由监督分类器给出；而文献[33]则假设一个可靠的预测结果，应该是在不同分类评价指标上都达到最优值；基于最大间隔假设，文献[34]提出了 S4VM 模型，它旨在从许多间隔较大的分界面中，通过优化无标签数据的标签分配（同时增加分界面的多样性），使得在满足间隔要求的条件下，寻找出多样性最大的分界面。这时，在极端条件下，存在一条分界面，满足间隔要求，但是它与其他分界面太过于不同，这种情况下，文章给出了 S4VM 性能的下线以及相关证明，说明 S4VM 的性能一定不会显著的比 只使用标签数据情况时的 SVM 差，所以它是安全的。但是，文献[34]的证明是基于一个较强的假设：真实的分界面是存在于已知分界面集合中，并且 S4VM 并不能处理多标签的情况。为了解决多标签分类问题，文献[64]是 S4VM 的多标签分类的扩展版本，它使用了层次的树形结构。首先使用半监督约束聚类，然后再得到的簇上，根据簇的中心距离建立二叉树，在最后从根节点开始，从上到下，在每个节点上执行 S4VM 算法。

从上文的分析中可以看出，基于集成框架的模型的最大问题，在于（1）要维护多个初始的半监督模型，最终的模型结果与初始模型数量和多样性有关；（2）只能线下的对多个初始模型结果进行可靠性分析，时效性太差；（3）对于大量数据来说，训练多个模型会是很耗时的操作。

（2） 基于半监督与监督方法的权衡模型

基于半监督和监督方法的权衡模型旨在，对于一个给定的监督模型，作者拟学习出一个半监督模型，并且约束每一个数据点在两个模型上的预测结果的一致性，保证这个半监督模型预测结果尽量不差于监督模型。其中基于 SSCCM 模型，文献[36]最小化监督模型和半监督模型在数据上的平方损失，并通过引入超参控制约束项的大小；文献[35]则基于 LapRLS 模型，同时也约束模型在数据上差异的平方损失，但每项约束强度大小由该数据的风险度给定，其中风险度由该数据的标签一致性和置信度给定。这种模型的不足之处仍在于需要额外维护一个监督模型，并且与半监督模型的学习过程有较强的依赖，如果在标签数据极少的情况下，该监督模型往往会欠拟合，进而严重影响半监督模型的准确性。

（3） 基于自适应权重模型

基于自适应权重的模型旨在对数据点学习出不同的权重，或者学习出数据之间的相似度权重，以此降低无关的或者不利于分类的无标签数据的权重，使得模型更加鲁棒可靠。文献[37]和文献[38]旨在对分界面附近的无标签数据点进行降权，拟达到最大化分类间隔的目的，但是这两个模型只能处理分割面附近的无标签数据，对于远离分割面的无标签数据都默认是安全可靠的；为了不仅仅关注于分割面附近的无标签数据，文献[47]基于网络中的节点中心性指标来衡量，一个无标签数据在多少个不同标签实例之间，以从没有帮助的无标签数据中选出可能有帮助的点，但前提需要知道哪些无标签数据是没有帮助的；而文献[43][44]也是关注于无标签数据的权重，通过密度估计和加权似然最大化的方式以实现安全半监督建模。文献[39]利用基于图的半监督模型，旨在学习数据之间的相似度权重，该模型不断迭代的降低具有不同标签的数据之间的权重，然后在新构造的图上进行重新训练半监督模型，直至收敛。

2、目前存在的问题

现有的半监督学习模型常常建立在对数据的强假设上，认为在同一个簇结构或流型结构上的数据拥有相似的标签。这样的假设加大了模型，假设和数据三者之间的强烈依赖，使得当前的半监督学习模型在部分数据集上的性能可能大大降低。可以从上文的分析中看出，现有的可靠性半监督学习算法还存在很多的不足，并且这些模型都没有考虑半监督数据流数据，而这都是亟待解决的问题：

第一，通过集成学习的方式就多个模型的输出进行经验性分析，最终模型效果与集成子模型的个数和多样性存在强烈依赖，也忽视了模型的实时性与简洁性；

第二，直接约束与监督模型的差异性，而忽略该监督模型本身训练误差，同时其约束力度也值得考究；

第三，通过建立数据点或数据点之间的权重学习模型，但缺乏深入探索数据固有的结构信息，而单独依赖特定分类器的划分效果，使得权重值可能与所选分类器有关，不能反映真实的权重信息。

第四，由于现有的可靠性模型建模的角度和模型的复杂性，使得它们并不能处理数据流数据，这在实际应用中还是一块很大的空缺。

所以，如何更加可靠有效的进行半监督学习，并且如何在半监督数据流上进行可靠有效的建模，安全的利用数据中的无标签数据来帮助模型的建立，提高模型的性能，成为了半监督学习理论和实践中的关键性问题。

【参考文献】

[1]. Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews][J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542-542.

[2]. Jebara T, Wang J, Chang S F. Graph construction and b-matching for semi-supervised learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 441-448.

[3]. Yan S, Wang H. Semi-supervised Learning by Sparse Representation[C]//SDM. 2009: 792-801.

[4]. Zhou D, Hofmann T, Schölkopf B. Semi-supervised learning on directed graphs[C]//Advances in neural information processing systems. 2004: 1633-1640

[5]. Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C]// Proceedings of the 20th Annual International Conference on Machine Learning. ACM. 2003, 3: 912-919

[6]. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. Journal of machine learning research, 2006, 7(Nov): 2399-2434.

[7]. Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[J]. Advances in neural information processing systems, 2004, 16(16): 321-328

[8]. Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs[C]//International Conference on Computational Learning Theory. Springer Berlin Heidelberg, 2004: 624-638.

[9]. Zhang Y M, Zhang X Y, Yuan X T, et al. Large-Scale Graph-Based Semi-Supervised Learning via Tree Laplacian Solver[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[10]. Wang F, Zhang C. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55-67.

[11]. Wu M, Schölkopf B. Transductive Classification via Local Learning Regularization[C]//AISTATS. 2007: 628-635.

[12]. Wang F, Li T, Wang G, et al. Semi-supervised Classification Using Local and Global Regularization[C]//AAAI. 2008, 8: 726-731

[13]. Bennett K, Demiriz A. Semi-supervised support vector machines[J]. Advances in Neural Information processing systems, 1999: 368-374.

[14]. Joachims T. Transductive inference for text classification using support vector machines[C]// Proceedings of the 16th Annual International Conference on Machine Learning. ACM. 1999, 99: 200-209.

[15]. Xu Z, Jin R, Zhu J, et al. Adaptive regularization for transductive support vector machine[C]//Advances in Neural Information Processing Systems. 2009: 2125-2133

[16]. Huang K, Xu Z, King I, et al. Supervised self-taught learning: Actively transferring knowledge from unlabeled data[C]//2009 International Joint Conference on Neural Networks. IEEE, 2009: 1272-1277

[17]. Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[18]. Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(6): 1088-1098.

[19]. He J, Carbonell J G, Liu Y. Graph-Based Semi-Supervised Learning as a Generative Model[C]//IJCAI. 2007, 7: 2492-2497.

[20]. Liu H, Yang Y. Semi-Supervised Learning with Adaptive Spectral Transform[C]//Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. 2016: 902-910

[21]. Goldberg A B, Li M, Zhu X. Online manifold regularization: A new learning setting and empirical study[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2008: 393-407.

[22]. Masud M M, Woolam C, Gao J, et al. Facing the reality of data stream classification: coping with scarcity of labeled data[J]. Knowledge and information systems, 2012, 33(1): 213-244.

[23]. Mallapragada P K, Jin R, Jain A K, et al. Semiboost: Boosting for semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(11): 2000-2014.

[24]. Wang F, Wang X, Li T. Beyond the graphs: Semi-parametric semi-supervised discriminant analysis[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 2113-2120.

[25]. Wang F, Zhang C. On Discriminative Semi-Supervised Classification[C]//AAAI. 2008: 720-725.

[26]. Cai D, He X, Han J. Semi-supervised discriminant analysis[C]//2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-7

[27]. Wang C, Yan S, Zhang L, et al. Non-Negative Semi-Supervised Learning[C]//AISTATS. 2009: 575-582

[28]. Li L, Zhao Z, Hou C, et al. Semi-supervised Learning Using Nonnegative Matrix Factorization and Harmonic Functions[M]//Computer Engineering and Networking. Springer International Publishing, 2014: 321-328

[29]. Wang F, Zhang C. Semisupervised learning based on generalized point charge models[J]. IEEE Transactions on Neural Networks, 2008, 19(7): 1307-1311.

[30]. Wang J, Shen X, Pan W. On efficient large margin semisupervised learning: Method and theory[J]. Journal of Machine Learning Research, 2009, 10(Mar): 719-742.

[31]. Li Y F, Zhou Z H. Improving semi-supervised support vector machines through unlabeled instances selection[J]. arXiv preprint arXiv:1005.1545, 2010.

[32]. Wang Y F L S B, Zhou Z H. Graph Quality Judgement: A Large Margin Expedition[J].

[33]. Li Y F, Kwok J T, Zhou Z H. Towards safe semi-supervised learning for multivariate performance measures[C]//Proceedings of 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ. 2016.

[34]. Li Y F, Zhou Z H. Towards making unlabeled data never hurt[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(1): 175-188.

[35]. Gan H, Luo Z, Sun Y, et al. Towards designing risk-based safe Laplacian Regularized Least Squares[J]. Expert Systems with Applications, 2016, 45: 1-7.

[36]. Wang Y, Chen S. Safety-aware semi-supervised classification[J]. IEEE transactions on neural networks and learning systems, 2013, 24(11): 1763-1772.

[37]. Huang K, Xu Z, King I, et al. Semi-supervised learning from general unlabeled data[C]//2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 273-282.

[38]. Wang D, Nie F, Huang H. Large-scale adaptive semi-supervised learning via unified inductive and transductive model[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 482-491.
[39]. Zhang Y M, Zhang Y, Yeung D Y, et al. Transductive Learning on Adaptive Graphs[C]//AAAI. 2010.
[40]. Singh A, Nowak R, Zhu X. Unlabeled data: Now it helps, now it doesn't[C]//Advances in neural information processing systems. 2009: 1513-1520.
[41]. M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In NIPS, pages 89–96, 2004
[42]. R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In ICML, pages 25–32, 2007
[43]. Kawakita M, Takeuchi J. Safe semi-supervised learning based on weighted likelihood[J]. Neural Networks, 2014, 53: 146-164.
[44]. Sokolovska N, Cappé O, Yvon F. The asymptotics of semi-supervised learning in discriminative probabilistic models[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 984-991.
[45]. Cozman F G, Cohen I, Cirelo M. Unlabeled Data Can Degrade Classification Performance of Generative Classifiers[C]//FLAIRS Conference. 2002: 327-331.
[46]. Krijthe J H, Loog M. Projected Estimators for Robust Semi-supervised Classification[J]. arXiv preprint arXiv:1602.07865, 2016
[47]. Chen S, Zhang C. Selecting Informative Universum Sample for Semi-Supervised Learning[C]//IJCAI. 2009, 6: 1016-1021
[48]. Wei L, Keogh E. Semi-supervised time series classification[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 748-753.
[49]. Chen Y, Hu B, Keogh E. Time Series Semi-Supervised Learning from a Single Example[C] //Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.
[50]. Yang Z, Cohen W, Salakhutdinov R. Revisiting Semi-Supervised Learning with Graph Embeddings[J]. arXiv preprint arXiv:1603.08861, 2016
[51]. Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines[J]. The Journal of Machine Learning Research, 2008, 9: 203-233.
[52]. Xu Z, Jin R, Zhu J, et al. Efficient convex relaxation for transductive support vector machine[C]//Advances in neural information processing systems. 2008: 1641-1648.
[53]. Nguyen T, Liu W, Perez E, et al. Semi-Supervised Learning with the Deep Rendering Mixture Model[J]. arXiv preprint arXiv:1612.01942, 2016.
[54]. Goldberg A B, Zhu X, Singh A, et al. Multi-Manifold Semi-Supervised Learning[C]//AISTATS. 2009: 169-176.
[55]. Wang F, Zhang C. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55-67.
[56]. Wang J, Wang F, Zhang C, et al. Linear neighborhood propagation and its applications[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(9): 1600-1615.
[57]. Huang K, Xu Z, King I, et al. Supervised self-taught learning: Actively transferring knowledge from unlabeled data[C]//2009 International Joint Conference on Neural Networks. IEEE, 2009: 1272-1277.
[58]. Azriel D, Brown L D, Sklar M, et al. Semi-Supervised linear regression[J]. arXiv preprint arXiv:1612.02391, 2016.
[59]. Alex V, Vaidhya K, Thirunavukkarasu S, et al. Semi-supervised Learning using Denoising Autoencoders for Brain Lesion Detection and Segmentation[J]. arXiv preprint arXiv:1611.08664, 2016.
[60]. Brefeld U, Scheffer T. Semi-supervised learning for structured output variables[C]//Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 145-152.
[61]. Brefeld U, Büscher C, Scheffer T. Multi-view discriminative sequential learning[C]//European Conference on Machine Learning. Springer Berlin Heidelberg, 2005: 60-71.
[62]. Sindhwani V, Niyogi P, Belkin M. A co-regularization approach to semi-supervised learning with multiple views[C]//Proceedings of ICML workshop on learning with multiple views. 2005: 74-79.

[63]. Leskes B. The value of agreement, a new boosting algorithm[C]//International Conference on Computational Learning Theory. Springer Berlin Heidelberg, 2005: 95-110.

[64]. Covões T F, Barros R C, da Silva T S, et al. Hierarchical bottom-up safe semi-supervised support vector machines for multi-class transductive learning[J]. Journal of Information and Data Management, 2013, 4(3): 357.

[65]. Yang T, Priebe C E. The effect of model misspecification on semi-supervised classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 33(10): 2093-2103.

[66]. Kääriäinen M. Generalization error bounds using unlabeled data[C]//International Conference on Computational Learning Theory. Springer Berlin Heidelberg, 2005: 127-142.

[67]. Rigollet P. Generalization error bounds in semi-supervised classification under the cluster assumption[J]. Journal of Machine Learning Research, 2007, 8(Jul): 1369-1392.

[68]. Bégin L, Germain P, Laviolette F, et al. PAC-Bayesian Theory for Transductive Learning[C]//AISTATS. 2014: 105-113.

[69]. Nigam K, Ghani R. Understanding the behavior of co-training[C]//Proceedings of KDD-2000 workshop on text mining. 2000: 15-17.

ABSTRACT

With the boom of the global information, while people are overwhelmed by the ocean of different kinds of data, technologies, such as data mining, artificial intelligence and machine learning, are playing a crucial role in the success of many applications, such as computer vision, natural language processing, deep learning and so on. However, the prevalence of data mining and machine learning in those fields requires that all the data that the model needs to train need to be fully labeled, namely, we need labeled data to train a supervised model. Unfortunately, real life data set are often unlabeled and fully labeled data are so hard to obtain, in this case, in order to exploit some valuable information from tons of irregular data, people have to personally label all the data instances and then perform some supervised algorithms that satisfy our tasks, which turns out to be very time-consuming and laborious. On the other hand, great changes have taken place in the way of data generation in our modern life. Specifically speaking, instead of static data, data in various fields are often presented in the form of stream, which makes people labeling data impossible to carry on, in fact, it will be an endless job.

Fortunately, the rise of semi-supervised learning overcomes this embarrass situation by allowing people to train a model with a small amount of labeled data and lots of unlabeled data. In other words, it targets to capitalize on the abundance of unlabeled data to obtain a better classifier than its counterpart supervised model which only uses labeled data to train. During the past decade, semi-supervised model has gained growing attentions in many applications such as classification, clustering, regression and data streaming mining. However, in spite of the popularity of semi-supervised technology in daily life, it still remains some problems waiting us to be solved, the most interesting problem among all is the reliability of semi-supervised model [40][45], which points to the situation that compared with its supervised model, the prediction performance of semi-supervised model, that utilizing the additional information of unlabeled instances in our data set, would be worse than its counterpart. Therefore, developing a better model that could more safely exploit the information of unlabeled data is a very important research, and worthy of the widespread commercial value and the prospects for development at the same time.

To this end, many approaches have been proposed from different points of view. In this report, by collecting some example algorithms, I introduce some common kinds of models of semi-supervised learning, including co-training model, max-margin model, graph-based model, online semi-supervised model and some other models that also act actively in some recent research areas.

However, Semi-supervised classification aims to generate a classifier of better performance together with plenty of unlabeled data. In most existing methods, however, known no other information on the distribution of those unlabeled data, unlabeled data are treated equally in many models without considering whether it is safe to use their additional message and some empirical evidence and theoretical characterization do [44] [45] indicate that unlabeled data help while in other situation, it doesn't. Since those unlabeled data introduce no any discriminative information on distribution of labels, semi-supervised model has to make certain assumptions (for example clustering assumption and manifold assumption) to work with unlabeled data, so any kind of assumption violation can lead to a slippery slope in classifier performance on both training data and test data. Therefore, to deal with this problem, designing a reliable and considerate model plays a crucial role in success of semi-supervised classification problem and toward safely using unlabeled instances to help semi-supervised leaning, I then introduce some reliable semi-supervised models, which can be mainly categorized into three classes, including ensemble-based model, semi-supervised and supervised trade-off model and adaptive weights model. Also several popular algorithms of different kinds are briefly introduced in this section and some of their shortcomings is concluded as well.

In the final part, I discuss and conclude the insufficient of existing reliable semi-supervised models in some detail. In general, it can be divided into three categories. Firstly, due to the analysis on the prediction output of every sub-models, the ensemble-based model’s performance is highly relied on the number and diversity of each sub-models. And the simplicity and real-time of sub-models is also ignored. Secondly, the semi-supervised and supervised trade-off model tries to minimize the difference of the prediction performance of the model using labeled data only and using all of data, which implicitly gives 100% trust on supervised classifier and is blind by the fact that only small number of labeled data is not sufficient to train any supervised model. On the other hand, the constraint effort of the prediction difference is still remain unanswered. Thirdly, the adaptive weights model learns the weights for each data or similarity measure without full exploration of the inherent structure of the data information. Depending on the classification effect of the specific classifier, the weight may be related to this selected classifier, which can’t reflect the real weight information of data. Finally, all those existing reliability models mentioned above in semi-supervised setting are unable to process the data stream, which still remains an insufficiency in the practical application. In fact, rather than considering the ensemble prediction or a single classifier output, one should pay more attention to the relation of each data points and the unsupervised feature information revealed by the inherent attribute of data structure.

To sum up, it’s a key problem but also an interesting challenge in semi-supervised learning theory and its applications that utilizes semi-supervised learning in a more reliable and safe way. And those problems are what I'm going to try to solve in my near future, including proposing a reliable semi-supervised classification model to adapt to the situation even when the labeled data are very rare, which is capable of avoiding these shortcomings via learning adaptive weights for each unlabeled instances and extending to handle with the evolving data stream in real environment.

Keywords: Reliable Semi-supervised Learning, Unlabeled Data, Data Stream