UM ESTUDO SOBRE SELEÇÃO, DIMENSIONALIDADE E ROTULAÇÃO DE AMOSTRAS EM APRENDIZADO DE MÁQUINA BASEADO EM INSTANCIAS

A STUDY ON SAMPLE SELECTION, DIMENSIONALITY, AND LABELING IN INSTANCE-BASED MACHINE LEARNING

Douglas Willian Rittono BARBOSA

douglas.rittono@gmail.com

Centro Universitário Padre Anchieta

Centro Universitário Campo Limpo Paulista

Resumo

A classificação é um dos problemas mais pesquisados na comunidade de mineração de dados. Preparar corretamente os dados que serão utilizados no treinamento dos chamados algoritmos supervisionados é uma das tarefas cruciais. O desbalanceamento das classes, falta de labels, a alta dimensionalidade dos conjuntos de treinamento podem prejudicar todo o processo de generalização, tanto em relação a precisão da classificação, quanto ao desempenho computacional. Este artigo apresenta uma revisão sistemática da literatura sobre seleção, dimensionalidade e rotulação de amostras em aprendizado de máquina baseado em instâncias. 27 estudos foram analisados de acordo com as abordagens utilizadas (ex.: Novos algoritmos). Os resultados apontam avanços na solução do problema e questões de pesquisas para a área de aprendizado de máquina e relacionadas.

Palavras-Chave

Conjunto de Dados; Dimensionalidade; Rotulação; Seleção de Amostras; Classificação; Aprendizado de Máquina.

Abstract

Classification is one of the most researched problems in the data mining community. Correctly preparing the data that will be used to train the so-called supervised algorithms is one of the crucial tasks. Class imbalance, lack of labels, and high dimensionality of training sets can harm the entire generalization process, both in relation to classification accuracy and computational performance. This paper presents a systematic review of the literature on sample selection, dimensionality, and labeling in instance-based machine learning. 27 studies were analyzed according to the approaches used (e.g., new algorithms). The results indicate advances in solving the problem and research questions for the area of machine learning and related fields.

Keywords

Dataset; Dimensionality; Labeling; Sample Selection; Classification; Machine Learning.

INTRODUÇÃO

O aprendizado de máquina tem desempenhado um papel importante no desenvolvimento de diferentes soluções. Os algoritmos de aprendizado supervisionado estão presentes principalmente nas tarefas de classificação. A função de tais algoritmos é prever o valor para qualquer objeto de entrada válido, após ter visto um número de exemplos de treinamento relacionados ao domínio. Muitas técnicas de aprendizado supervisionado visam a classificação binária (PASSERINI, PONTIL & FRANSCONI, 2004). No entanto, muitos problemas reais lidam com a chamada classificação multi-classe, que consistem em instancias que possuem duas ou mais classes (FARID et al. 2014). Existem ainda as instâncias multi-label, que lidam com N valores de classe, cada grupo de valores de classe contém um conjunto de instancias de dados com N características de entrada.

Quando um conjunto de dados de treinamento possui a maioria das amostras em algumas classes, enquanto a minoria pertence a outras, diz-se que os dados são desbalanceados (JAPKOWICZ 2000). Como os classificadores tradicionais tendem a se concentrar nas classes majoritárias, as classes minoritárias acabam perdidas nas previsões, e com isso, tem um desempenho ruim.

A preparação de dados não é uma tarefa trivial e depende da natureza dos dados. Existem muitos problemas que precisam ser abordados durante a preparação de dados, como a existência de outliers, erros, ruídos, valores ausentes, etc. Neste contexto este artigo tem como principal objetivo responder a seguinte questão de pesquisa: “Como preparar as instancias de dados de maneira eficiente para melhorar os classificadores, em cenários binário, multi-classe e multi-label?”.

Os principais objetivos desta revisão são: (1) identificar as técnicas e teorias utilizadas, os aspectos positivos e limitações dos estudos existentes e (2) apontar lacunas na literatura e desafios de pesquisa, atuais e futuros. O restante deste artigo está organizado da seguinte maneira: em **Trabalhos Relacionados** é apresentada uma revisão sobre o tema relacionado a questão principal, porém, aplicado a um domínio de negócio especifico; em **Metodologia da Revisão** está descrita a metodologia utilizada para a realização da revisão e análise dos artigos; em **Resultados da Revisão** há um detalhamento dos resultados obtidos e categorizados pelas técnicas empregadas; e em **Discussão e Conclusão** é apresentada uma discussão sobre os resultados obtidos e as conclusões.

TRABALHOS RELACIONADOS

Após ajuste na busca, uma revisão sistemática foi identificada. O trabalho foi enquadrado de acordo com seus objetivos e aplicações. Uma análise sintética desse trabalho é apresentada a seguir.

BUJANG et al. (2023), apresentam uma revisão sobre as abordagens existente na tratativa de dados desbalanceados, binários e multi-classe, aplicados a classificação no ensino superior. Segundo os autores, uma das maneiras de medir o sucesso dos estudando é prever seu desempenho com base em suas notas acadêmicas anteriores. Argumentam também que vários modelos preditivos são amplamente desenvolvidos e aplicados nesse sentido, porém, sem focar em dados desbalanceados. Apresentam em seu estudo os métodos de balanceamento mais comuns publicados entre 2015 e 2021 e destacam seu impacto na resolução em três abordagens, nível de dados, nível de algoritmo e nível híbrido.

A revisão apresentada neste artigo se diferencia da analisada nos seguintes aspectos: (1) o foco desta revisão está no tratamento dos dados baseados em instancias independente de domínio de negócio, enquanto a revisão citada pretende avaliar para um domínio especifico, a classificação para o ensino superior; (2) a revisão analisada não faz uma análise sobre dados multi-label, somente multi-classe, enquanto esta investiga o tratamento dos conjuntos de dados de uma forma mais abrangente.

METODOLOGIA DA REVISÃO

A revisão sistemática da literatura apresentada neste artigo teve como base metodológica o guia apresentado em (KITCHENHAM, 2004). Este estudo tem como principal objetivo responder a seguinte questão de pesquisa: “Quais são as abordagens para reduzir amostras, reduzir dimensionalidade ou rotular os dados utilizados em aprendizado de máquina baseado em instancias?”.

Uma pesquisa exploratória preliminar baseada na questão de pesquisa foi realizada com o objetivo de levantar insumos necessários à pesquisa, resultando na definição dos parâmetros da pesquisa, no período de abrangência da busca, nas bases científicas e palavras-chave a serem utilizadas, e na área de busca nos artigos. O período de busca (2009 a 2023) se deve ao fato de ser um tema que já vem sendo estudados a alguns anos, bem como se espera relatar os avanços nos últimos anos. A seguinte string de busca foi utilizada: “(Machine Learning) AND (Instance-Based OR Instance Based) AND (Multiclass OR Multi-class OR Multi-label OR Multilabel) AND (Reduce OR Reducing)”. A execução da busca nas bases científicas considerou todos os artigos retornados. Assim, a busca inicial obteve um total de 73 artigos, todos da base IEEE Xplore.

Os critérios de inclusão e exclusão foram definidos por um pesquisador, especialista da área da computação, em um processo iterativo de leitura de artigos (na busca exploratória) e proposição de critérios até atingir consenso. Os critérios estão detalhados na Tabela 1. A primeira avaliação considerou o título, resumo e palavras-chave.

Tabela 1 – Critérios de inclusão e exclusão de artigos

Tipo	Sigla	Critério
Inclusão	I1	Pesquisas sobre seleção de amostras.
	I2	Pesquisas sobre redução de dimensionalidade.

	I3	Pesquisas sobre rotulação de amostras faltantes.
	I4	Estudos que abordam dados Multi-classe e Multi-label.
Exclusão	E1	Artigos escritos em idiomas diferentes do Inglês e do Português.
	E2	Artigos que não estejam relacionados com aprendizado de máquina baseado em instancias.
	E3	Artigos que não sejam da área de computação ou multidisciplinar com computação.
	E4	Textos que não sejam publicações científicas.
	E5	Resumos com menos de 4 páginas e que não tenham profundidade ou resultados relevantes.
	E6	Revisões sistemáticas e Livros.

Os 27 artigos categorizados como trabalhos com possibilidade de aderência ao tema da pesquisa foram avaliados em sua totalidade perante os critérios. Na busca utilizada não foram encontrados trabalhos de revisão de literatura relacionadas ao tema, portanto, foi necessário adaptar a busca, com isso 1 trabalho foi identificado como trabalho relacionado.

RESULTADOS DA REVISÃO

Esta seção apresenta uma análise sintética dos 29 estudos selecionados. A primeira Subseção apresenta soluções baseadas em novos algoritmos, enquanto a segunda apresenta soluções baseadas em propostas de modelos, frameworks ou avaliação dos modelos existentes. A Tabela 2 apresenta uma síntese dos trabalhos analisados.

Tabela 2 – Síntese dos trabalhos analisados

(continua)

Autores	*Objetivos*				*Características*									
	C	**D**	**R**	**P**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
(Gong & Zhai., 2021)	X				X									
(Pham et al., 2017)	X	X					X				X	X		
(Yu et al., 2021)	X			X			X		X					
(Kumari & Thakar., 2017)	X		X				X	X				X		X
(Davuluri et al., 2023)	X			X	X			X	X	X			X	
(Ranganathan et al., 2012)	X			X				X	X				X	

(Ghosh & Bandyopadhyay., 2015)	X						X				X	X		X
(Abdi & Hashemi., 2016)	X		X				X	X				X		
(Shrivastava et al., 2014)	X					X					X			
(Wang & Xu, 2023)				X					X	X				
(Du et al., 2017)			X	X	X				X			X		
(Zdravevski et al., 2015)			X	X			X	X	X	X				
(Adhikari et al., 2018)	X	X												
(Pervez & Farid., 2014)	X	X					X			X			X	
(Rocha & Goldenstein., 2014)	X						X			X		X		
(Mahfuzh & Purwarianti., 2022)			X	X					X	X				
(Hussien et al., 2021)		X		X						X				
(Sheikh-Nia et al., 2012)	X		X				X	X						
(Huang et al., 2014)			X		X									
(Firouzi et al., 2017)			X				X			X				
(Raja & Arunadevi., 2023)	X		X			X		X					X	

Tabela 2 – Síntese dos trabalhos analisados

(conclusão)

Autores	*Objetivos*				*Características*									
	C	**D**	**R**	**P**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
(Kumar et al., 2016)	X		X				X						X	
(Adbulhammed et al., 2019)	X	X								X		X		
(Sudharson et al., 2021)	X							X	X					
(Chakraborty et al., 2015)			X		X									
(Af'Idah et al., 2023)	X							X	X					
(Gao et al., 2018)	X			X			X			X				X
(Xu & Xu., 2017)	X			X			X			X		X		

Legenda: *Objetivos*: (C) Classificação; (D) Detecção; (R) Rotulação; e (P) Performance. *Características*: (1) Novo Algoritmo; (2) Novo Framework; (3) Novo Método; (4) Dados Desbalanceados; (5) Redução de Tempo; (6) Redução de Dimensionalidade; (7) Sacola de Instancias; (8) Calculo de Distância; (9) SVM; (10) k-NN.

Soluções baseadas em Novos Algoritmos

GONG & ZHAI (2021), DAVULURI, SRIVASTAVA, AERI, ARORA, KESHTA & RIVERA (2023), DU, WANG, ZHANG, ZHANG & TAO (2017), HUANG, JIN, ZHOU (2014) e CHAKRABORTY, BALASUBRAMANIAN, SUN, PANCHANATHAN & YE (2015) propõem novos algoritmos para tratar as amostras, além de demonstrar a eficácia das novas possibilidades.

GONG & ZHAI, (2021) alegam que existem poucos estudos para classificação de dados multi-label, dessa forma, propõem um novo algoritmo de classificação ativa online, baseada em uma estratégia hibrida de consulta, durante o experimento, analisam o novo algoritmo aplicado a 6 datasets multi-label diferentes, demonstrando a eficiência da proposta.

Segundo DAVULURI et al. (2023) o desequilíbrio na distribuição das amostras entre categorias é um ponto crucial na precisão e performance dos modelos atuais, portanto, criam um novo algoritmo denominado MOIS, tal algoritmo, toma o centro do cluster como ponto de referência, enquanto remove amostras de treinamento redundantes e seleciona as amostras de fronteira decisivas para reduzir significativamente os dados de treinamento, após esse processo, é possível obter um bom desempenho na classificação de dados multi-classe mesmo com grande volume de amostras usando SVM (Support Vector Machine).

Em DU et al. (2017) temos a criação de uma derivação de um algoritmo de aprendizado de máquina multi-label baseado em MCC (Maximum Correntropy Criterion), que visa trazer uma fusão de incerteza e representatividade com os labels de previsão para dados desconhecidos, o intuito do trabalho é reduzir os custos de rotulagem e melhorar a capacidade de treinar um bom modelo de multi-label simultaneamente.

HUANG, JIN & ZHOU, (2014) abordam a limitação dos algoritmos Active Learning ao usar instancias não rotuladas que são informativas ou representativas e propõem então uma abordagem baseada na visão min-max estendendo o algoritmo QUIRE. Incorporam a correlação entre labels para aprendizado multi-label ao solicitar ativamente pares de instance-label. Os resultados experimentais demonstram que a abordagem proposta supera várias abordagens de Active Learning estado-da-arte, tanto em aprendizado multi-label, objeto deste estudo, quanto em aprendizado de single-label.

Na mesma direção CHAKRABORTY et al. (2015) apresentam dois novos algoritmos de Active Learning, porém, utilizando o modo batch (BMAL), chamados BatchRank e BatchRand, a intenção é rotular dados não rotulados com base nos já existentes. Utilizando 15 conjunto de dados binários, multi-classe e multi-label, demonstram que os algoritmos propostos têm desempenho comparável às técnicas de ponta, entregam soluções de alta qualidade e são robustos a problemas do mundo real com ruído nos rótulos e desequilíbrio de classes.

Soluções baseadas em Modelos, Frameworks ou Avaliação de modelos existentes

Dentre os 27 estudos avaliados, 22 exploram modelos, frameworks ou avaliam técnicas já existentes, indo em uma direção diferente das propostas de novos algoritmos vistos na sessão

anterior. Os estudos em questão, abordam em sua maioria, questões de redução de dimensionalidade, redução de tempo de execução e dados desbalanceados. A seguir apresenta-se uma síntese dos trabalhos.

PHAM, RAICH & FERN, (2017) propõem um modelo probabilístico discriminativo para o problema de anotação de instancias, buscam com tal modelo, inferir labels de dados nas chamadas, bolsas de instancias, de maneira performática. Introduzem também uma estrutura de maximização da expectativa para inferência, baseada na abordagem de máxima verossimilhança.

Utilizando as bolsas de instancias GHOSH & BANDYPPADHYAY, (2015) introduzem uma técnica baseada em Citation-KNN Fuzzy, utilizando a distância de Hausdorff, dessa forma, trazem uma redução do efeito de instancias falso-positivo em sacolas positivas. Em seus experimentos em conjuntos de dados de descoberta de medicamentos e imagens, verificam um melhor desempenho que o tradicional Citation-KNN e competitivo com a maioria dos algoritmos de última geração.

Ainda trabalhando com bolsas de instancias SHRIVASTAVA et al. (2014) apresentam um framework de otimização baseado no modelo noisy-OR para o aprendizado dos dicionários. Vários experimentos usando os conjuntos de dados multi-classe populares, mostram que o método proposto apresenta um desempenho comparável aos métodos existentes.

Ao fazer uso do modelo BERT, estado da arte, como modelo base (YU et al., 2021) constroem um modelo multitarefa que possa compartilhar bem os recursos no aprendizado. Uma característica importante no trabalho é o uso de dados multi-label hierárquicos, portanto, os autores buscam também entender como combinar os labels previstos de diferentes níveis na hierarquia.

Em KUMARI & THAKAR, (2017) o conjunto de dados desequilibrado é tratado com a criação de dados sintéticos usando um método de sobre amostragem baseado na distância de Hellinger, equilibrando assim a amostragem. Após o balanceamento dos dados o conjunto é submetido a classificação utilizando algoritmos clássicos como k-NN e C4.5, os resultados demonstram um aumento de 20% na precisão em comparação com a classificação do mesmo conjunto de dados sem o tratamento proposto.

ABDI & HASHEMI, (2015) tratam do desequilíbrio dos dados de forma similar, utilizam a criação de dados sintéticos com um método de sobre amostragem baseado na distância de Mahalanobis. Dessa vez os experimentos utilizam como base de comparação a métrica MAUC (Mean Area Under the Curve) e precisão sobre outros modelos populares de sobre amostragem, a analise teórica e as observações empíricas, demonstram uma melhoria em alguns benchmarks.

RANGANATHAN, RAMANAN & NIRANJAN, (2012) buscam otimizar a classificação de dados multi-label com SVM. Nesse sentido propõem uma nova arquitetura que denominam como UDT (Unbalanced Decision Tree), com isso, reduziram drasticamente o tempo de treinamento, encontrando a ordem dos classificadores com base em seus desempenhos durante a seleção de nó raiz e fixando essa ordem para formar a hierarquia da árvore de decisão. O uso da UDT com SVM envolve menos classificadores do que OVO, OVA e DAGSVMs, mantendo uma precisão comparável a essas técnicas padrão, portanto, há uma redução de tempo de treinamento em comparação com as arquiteturas já existentes.

PERVEZ & FARID, (2014) utilizam do SVM para propor uma nova abordagem de detecção e classificação de intrusão em sistemas de rede baseado em computadores. O uso do SVM em conjunto com uma seleção de características para o conjunto de dados de DI multi-classe "NSL-KDD CUP 99" demonstraram que a seleção eficaz de características pode reduzir significativamente o número de dimensões de entrada sem sacrificar a precisão da classificação, o que melhora a eficiência computacional e a escalabilidade dos sistemas de detecção de intrusão.

Em KUMAR et al. (2016), busca-se classificar automaticamente web spam por tipo, para tal, os autores introduzem novas características baseadas em cloaking que ajuda o modelo a alcançar alta precisão e taxa de recall, reduzindo assim as taxas de falsos positivos. A abordagem de classificador denominada DMMH-SVM (Dual-Margin Multi-Class Hypersphere Support Vector Machine) tem sua eficácia justificada analiticamente no estudo, os resultados experimentais demonstram que o DMMH-SVM supera os algoritmos existentes com novas características de cloaking.

RAJA & ARUNADEVI, (2023) apresentam um framework de aprendizado ativo para análise de sentimentos multi-classe, focando no desequilíbrio de classe. Combinam CNN (Convolutional Neural Network), Regressão Logística, Random Forest, SVM e Gradiente Boosting para selecionar de forma inteligente instancias informativas para rotulação, reduzindo a necessidade de grandes quantidades de dados rotulados e esforços de anotação manual.

WANG & XU, (2022) buscam melhorar o desempenho na classificação de dados multi-label com RSVM (Rank Support Vector Machine). Os autores alegam que essa é a primeira tentativa de construir uma nova regra de triagem segura (SSR) para problemas de aprendizagem multi-label. O SSR pode filtrar e excluir a maioria das instancias com base em seus pares de labels relevantes-irrelevantes, após esse processo, a escala do RSVM pode ser substancialmente reduzida. Os extensos experimentos com cinco conjuntos de dados benchmark, três conjuntos de dados em larga escala e um conjunto de dados de diabetes do tipo 2 mostram a eficiência e segurança da abordagem.

Observa-se em ROCHA & GOLDENSTEIN, (2013) um esforço para encontrar os classificadores binários mais discriminativos para resolver problemas multi-classe e manter a eficiência. Os autores introduzem um conceito de correlação e probabilidade conjunta para conjuntos de larga escala. Utilizando de uma abordagem bayesiana, os autores buscam uma estratégia para reduzir o número de aprendizes base e outra estratégia para encontrar novos aprendizes base que possam complementar melhor o conjunto existente. Os experimentos validaram e compararam o método com um conjunto diversificado de métodos da literatura em vários conjuntos de dados públicos, e variam de problemas pequenos (10 a 26 classes) e problemas multiclasse grandes (até 1000 classes).

GAO et al., (2018) buscam resolver o problema de multi-label em larga escala propondo três métodos baseados em ML-KNN. O primeiro método PML-KNN se ampara no conceito de análise de componentes principais aplicados ao KNN, no segundo, além do citado anteriormente, os autores adicionam a similaridade acoplada, denominando PCSML-KNN, por último, unem a similaridade acoplada e seleção de características FCSML-KNN. Foram testados dados de dois conjuntos reais, nos três métodos propostos e no método base (ML-KNN), os resultados demonstram que a redução de dimensões dos métodos propostos pode melhorar a eficiência de classificação.

ABDULHAMMED et al., (2019) fazem uso de PCA para redução de dimensionalidade na detecção e classificação de intrusão de rede com dados desbalanceados. Na proposto os autores utilizam uma junção das técnicas Random Forest, Rede Bayesiana, Analise discriminante Linear e Análise discriminante Quadrática. Os achados experimentais foram capazes de reduzir as dimensões do conjunto de dados CICIDS2017 de 81 para 10 características e mostram melhor desempenho em termos de Taxa de Detecção, F-Measure, Taxa de Alarme Falso e Acurácia.

SHEIKH-NIA, GREWAL E AREIBI, (2012) propõem uma técnica que denominam SEC (Sequential Ensemble Classification) para reduzir o desequilíbrio de instancias transformando um problema multi-classe em uma sequência de problemas de classificação binária. No trabalho, os autores, fazem uma investigação de duas implementações diferentes do método proposto, uma baseada em um conjunto homogêneo de classificadores e outra em um conjunto heterogêneo. Para validação empírica, selecionam um conjunto de dados médicos do mundo real caracterizado por desbalanceamento significativo. Os resultados experimentais demonstram que ambas as versões do método SEC superam os classificadores individuais. Especificamente, o design homogêneo alcança o melhor desempenho entre as abordagens avaliadas.

Em AF'IDAH et al., (2023) é feita uma comparação de três técnicas para tratar dados desbalanceados na avaliação de sentimentos. Os resultados do estudo indicam que a técnica SMOTE é mais adequada do que ADSYN para classificações de aspecto multi-label que empregam LSTM ou BILSTM. Quando comparam a abordagem de sobre amostragem ao uso do SMOTE ou ADASYN, todos os escores de avalição em LSTM e BILSTM foram reduzidos. A conclusão é que o melhor desempenho do modelo para classificação de aspectos multi-label, é LSTM com implementação de SMOTE para fazer o balanceamento dos dados.

MAHFUZH & PURWARIANTI, (2022) buscam classificar textos multi-classe de forma performática. Com o intuito de reduzir o consumo de recursos através do compartilhamento de parâmetros e aumentar o desempenho obtendo características de outros labels durante o treinamento, portanto, constroem uma implementação de aprendizado multitarefa durante o ajuste fino de transformes em dados de relevância binária. Nos experimentos utilizam 1599 instancias de dados de treinamento e 400 instancias de dados de teste com 16 classes, o método proposto alcançou 0,8817 para classificação e 0,9083 em F1 para a classificação de 3 classes de amostragem, demonstrando, uma superação dos valores alcançados com ajuste fino multi-label padrão.

FIROUZI, KARIMIAN & SOLEYMANI, (2017) buscando reduzir a dimensionalidade do espaço de rótulos, propõem um método modificado de NMF (Fatoração de Matriz não Negativa). No estudo consideram que a matriz de labels é binária e que nesta matriz alguns labels importantes para uma instancia podem não estar presentes, chamados de labels ausentes. Os experimentos compararam os resultados com métodos de classificação multi-label estado da arte, e demonstram uma superioridade do método proposto.

XU & XU, (2017) trazem uma proposta de método de transformação de comparação par a par de labels (PCT) para fazer a seleção de características, visando uma melhoria no desempenho para classificação multi-label. Quatro conjuntos de dados textuais foram usados para o experimento. Os resultados mostram que a proposta supera cinco técnicas existentes de seleção de características do tipo filtro, baseada em transformação de acordo com seis medidas de avaliação.

Em ZDRAVEVSKI et al., (2015) a proposta é transformar dados nominais em numéricos se baseando em WOE (Weight of Evidence Parameter). Dessa forma, um número menor de

características é gerado em comparação a técnicas baseadas na geração fictícia de características, o que melhora a precisão da classificação, reduz a complexidade de memória e encurta o tempo de execução. Os autores apontam algumas fraquezas do método e fazem algumas recomendações de qual cenário utilizar um método ou outro.

Em seu trabalho TAO & GUIYANG, (2015) argumentam que o modelo BR (Binary Relevance) tem sido marginalizado na literatura devido a percepção de inadequação das correlações de labels. Então apresentam redes de dependência condicional aprimorada, e fazem a descrição de vários métodos de classificação baseados em BR. O modelo aplicado em conjuntos de dados de benchmark, demonstram que o mesmo obtém melhor desempenho preditivo em vários conjuntos de dados, sob vários métodos de avaliação especificamente projetados para classificação multi-label.

SUDHARSON et al., (2021) avaliaram o desempenho do framework AdaBoost melhorado para classificação multi-classe em conjuntos de dados desbalanceados. Os autores fizeram uma melhoria em fases randomizadas através de stumps. Ao aproveitar sua capacidade de ajustar pesos iterativamente e focar em instancias classificadas incorretamente, o AdaBoost pode melhorar a precisão da classificação, mesmo na presença de dados desbalanceados em cenários multi-classe.

DISCUSSÃO E CONCLUSÃO

Esta revisão sobre técnicas de preparação de dados em aprendizado de máquina revelou insights significativos sobre a importância da qualidade dos dados na performance dos classificadores. A análise dos estudos revisados destacou duas principais categorias de abordagens: novos algoritmos desenvolvidos para lidar com desafios específicos de dados e a adaptação de modelos, frameworks e técnicas existentes para melhorar a eficácia e eficiência dos sistemas de aprendizado de máquina.

Os estudos revisados mostram que a preparação adequada dos dados, incluindo o tratamento de desequilíbrios de classes, redução de dimensionalidade e criação de dados sintéticos, desempenha um papel crucial na obtenção de resultados precisos e robustos. A utilização de técnicas avançadas como Active Learning (GONG & ZHAI, 2021; CHAKRABORTY et al., 2015), métodos baseados em SVM (DAVULURI et al., 2023; RANGANATHAN et al., 2012), redes neurais convolucionais (RAJA & ARUNADEVI, 2023), e técnicas de sobreamostragem como SMOTE (AF'IDAH et al., 2023) demonstrou melhorias significativas na precisão e na capacidade de generalização dos modelos.

Em síntese, esta revisão sublinha a necessidade contínua de desenvolver e aprimorar técnicas de preparação de dados que sejam adaptáveis a diferentes contextos e desafios de aprendizado de máquina. Ao enfrentar eficazmente problemas como desequilíbrio de classes e alta dimensionalidade, os avanços na preparação de dados não apenas melhoram a precisão dos modelos, mas também aumentam sua aplicabilidade em diversas áreas, desde a detecção de intrusões em redes até a classificação de sentimentos em textos.

Os resultados indicam que futuras pesquisas devem se concentrar na validação e na aplicação prática dessas técnicas em cenários reais, garantindo que os benefícios teóricos se traduzam em melhorias tangíveis na eficiência e na confiabilidade dos sistemas de aprendizado de máquina.

Ao fazer isso, podemos avançar na fronteira do conhecimento em preparação de dados e promover inovações significativas em inteligência artificial e ciência de dados.

Este artigo apresentou uma revisão sistemática da literatura sobre seleção, dimensionalidade e rotulação de amostras em aprendizado de máquina baseado em instancias. A abordagem empregada nesta revisão da literatura permitiu a verificação e análise de tendências, bem como abordagens tecnológicas adotadas ao longo dos últimos quatorze anos. Este estudo se diferencia dos demais por sua abordagem abrangente, não restrita a um domínio de negócio específico, e por incluir uma análise de dados multi-label, além de dados binários e multi-classe. Isso proporcionou uma visão mais ampla das técnicas de preparação de dados e suas aplicações. Em um universo de 73 artigos inicialmente recuperados, 27 trabalhos foram criteriosamente selecionados, classificados e sintetizados de modo a representar o estado-da-arte. Foram apresentadas as técnicas e teorias utilizadas, os aspectos negativos e positivos de cada um e limitações dos estudos, também aponta lacunas na literatura e desafios de pesquisa, atuais e futuros.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDI, L., and HASHEMI, S. To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge and Data Engineering 28, 1 (2016), 238–251.

ABDUL Bujang, S. D., SELAMAT, A., KREJCAR, O., MOHAMED, F., CHENG, L. K., CHIU, P. C., and FUJITA, H. Imbalanced classification methods for student grade prediction: A systematic literature review. IEEE Access 11 (2023), 1970– 1989.

ABDULHAMMED, R., FAEZIPOUR, M., MUSAFER, H., and ABUZNEID, A. Efficient network intrusion detection using pca-based dimensionality reduction of features. In 2019 International Symposium on Networks, Computers and Communications (ISNCC) (2019), pp. 1–6.

AF'IDAH, D. I., ANGGRAENI, P. D., HANDAYANI, S. F., and DAIROH. Imbalanced classes treatment in deep learning multi-label aspect classification using oversampling and under-sampling. In 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE) (2023), pp. 755–760.

CHAKRABORTY, S., BALASUBRAMANIAN, V., Sun, Q., PANCHANATHAN, S., and YE, J. Active batch selection via convex relaxations with guaranteed solution bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 37, 10 (2015), 1945–1958.

DAVULURI, S. K., SRIVASTAVA, D., AERI, M., ARORA, M., KESHTA, I., and RIVERA, R. Support vector machine based multi-class classification for oriented instance selection. In 2023 International Conference on Inventive Computation Technologies (ICICT) (2023), pp. 112–117.

DU, B., WANG, Z., ZHANG, L., ZHANG, L., and TAO, D. Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion. IEEE Transactions on Image Processing 26, 4 (2017), 1694–1707. 6

FARID, D. M., ZHANG, L., RAHMAN, C. M., HOSSAIN, M., and STRACHAN, R. Hybrid decision tree and naive bayes classifiers for multi-class classification tasks. Expert Systems with Applications 41, 4, Part 2 (2014), 1937–1946.

FIROUZI, M., KARIMIAN, M., and SOLEYMANI, M. Nmf-based label space factorization for multi-label classification. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (2017), pp. 297–303.

GAO, S., YANG, X., ZHOU, L., and YAO, S. The research of multi-label k-nearest neighbor based on descending dimension. In 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA) (2018), pp. 129–135.

GHOSH, D., and BANDYOPADHYAY, S. A fuzzy citation-knn algorithm for multiple instance learning. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2015), pp. 1–8.

GONG, K., and ZHAI, T. An online active multi-label classification algorithm based on a hybrid label query strategy. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (2021), pp. 463–468.

HUANG, S.-J., JIN, R., and ZHOU, Z.-H. Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 10 (2014), 1936–1949.

JAPKOWICZ, N. The class imbalance problem: Significance and strategies. Proceedings of the 2000 International Conference on Artificial Intelligence ICAI (06 2000).

KITCHENHAM, B. Procedures for performing systematic reviews. Keele, UK, Keele Univ. 33 (08 2004).

KUMAR, S., GAO, X., WELCH, I., and MANSOORI, M. A machine learning based web spam filtering approach. In 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA) (2016), pp. 973–980.

KUMARI, A., and THAKAR, U. Hellinger distance based oversampling method to solve multi-class imbalance problem. In 2017 7th International Conference on Communication Systems and Network Technologies (CSNT) (2017), pp. 137–141.

MAHFUZH, M., and PURWARIANTI, A. Multi-label classification of Indonesian financial risk news using transformer-based multi-task learning. In 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA) (2022), pp. 1–6.

PASSERINI, A., PONTIL, M., and FRASCONI, P. New results on error correcting output codes of kernel machines. IEEE Transactions on Neural Networks 15, 1 (2004), 45–54.

PERVEZ, M. S., and FARID, D. M. Feature selection and intrusion classification in nsl-kdd cup 99 dataset employing svms. In The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014) (2014), pp. 1–6.

PHAM, A. T., RAICH, R., and FERN, X. Z. Dynamic programming for instance annotation in multi-instance multi-label learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 12 (2017), 2381–2394. 7

RAJA, M., and ARUNADEVI, J. Deep active learning multiclass classifier for the sentimental analysis in imbalanced unstructured text data. In 2023 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI) (2023), pp. 1–6.

RANGANATHAN, P., RAMANAN, A., and NIRANJAN, M. An efficient and speeded-up tree for multi-class classification. In 2012 IEEE 6th International Conference on Information and Automation for Sustainability (2012), pp. 190–193.

ROCHA, A., and GOLDENSTEIN, S. K. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. IEEE Transactions on Neural Networks and Learning Systems 25, 2 (2014), 289–302.

SHEIKH-NIA, S., GREWAL, G., and AREIBI, S. A sequential ensemble classification (sec) system for tackling the problem of unbalance learning: A case study. In 2012 11th International Conference on Machine Learning and Applications (2012), vol. 2, pp. 72–77.

SHRIVASTAVA, A., PILLAI, J. K., PATEL, V. M., and CHELLAPPA, R. Dictionary-based multiple instance learning. In 2014 IEEE International Conference on Image Processing (ICIP) (2014), pp. 160–164.

SUDHARSON, D., ASHFIA Fathima, S., KAILAS, P. S., THRISHA Vaishnavi, K. S., DARSHANA, S., and BHUVANESHWARAN, A. Performance evaluation of improved adaboost framework in randomized phases through stumps. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (2021), pp. 1–6.

TAO, G., and GUIYANG, L. Improved conditional dependency networks for multi-label classification. In 2015 Seventh International Conference on Measuring Technology and Mechatronics Automation (2015), pp. 561–565.

WANG, X., and XU, Y. Label pair of instances-based safe screening for multilabel rank support vector machine. IEEE Transactions on Systems, Man, and Cybernetics: Systems 53, 3 (2023), 1907–1919.

XU, H., and XU, L. Multi-label feature selection algorithm based on label pairwise ranking comparison transformation. In 2017 International Joint Conference on Neural Networks (IJCNN) (2017), pp. 1210–1217.

YU, Y., SUN, Z., SUN, C., and LIU, W. Hierarchical multilabel text classification via multitask learning. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (2021), pp. 1138–1143.

ZDRAVEVSKI, E., LAMESKI, P., KULAKOV, A., and KALAJDZISKI, S. Transformation of nominal features into numeric in supervised multi-class problems based on the weight of evidence parameter. In 2015 Federated Conference on Computer Science and Information Systems (FedCSIS) (2015), pp. 169–179.

AGRADECIMENTOS

Agradeço imensamente ao Prof. Dr. Ferrucio de Franco Rosa do Centro Universitário Campo Limpo Paulista que me deu todo o apoio e base necessários para elaboração deste trabalho. Muito obrigado Prof. Dr. Ferrucio, de Franco Rosa por seu compromisso e por compartilhar todo seu conhecimento.