Autoreferat dotyczący osiągnięć naukowych

1 Imię i nazwisko: Krzysztof Dembczyński

2 Posiadane dyplomy oraz stopnie naukowe

1. Dyplom doktora nauk technicznych w zakresie informatyka, Wydział Informatyki i Zarządzania, Politechnika Poznańska, 31 marca, 2009, tytuł rozprawy: *Decision Rule Model for Ordinal Classification Problems with Incomplete Information/Model regułowy dla problemów klasyfikacji porządkowej z informacją niepełną*, Promotor: prof. Roman Słowiński
2. Magister informatyki, Specjalność: Inteligentne systemy wspomagania decyzji, Wydział Elektryczny, Politechnika Poznańska, 2001
3. Inżynier informatyki, Wydział Elektryczny, Politechnika Poznańska, 1999

3 Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych

01.10.2009–obecnie	–	Adiunkt, Instytut Informatyki, Politechnika Poznańska
23.09.2009–22.09.2011	–	Pracownik naukowy, Wydział Matematyki i Informatyki, Uniwersytet w Marburgu
01.10.2001–30.09.2009	–	Asystent, Instytut Informatyki, Politechnika Poznańska

4 Wskazanie osiągnięcia naukowego

Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2016 r. poz. 882 ze zm. w Dz. U. z 2016 r. poz. 1311.):

4.1 Tytuł osiągnięcia naukowego: Algorytmy uczenia maszynowego dla problemów klasyfikacji wieloetykietowej

Moim głównym osiągnięciem naukowym są wyniki uzyskane w dziedzinie uczenia maszynowego, dotyczące problemu klasyfikacji wieloetykietowej. Zostały one opublikowane w cyklu artykuł naukowych na prestiżowych konferencjach uczenia maszynowego i sztucznej inteligencji, takich jak *International Conference on Machine Learning*, *Neural Information Processing Systems* i *European Conference on Artificial Intelligence*, oraz w uznanych czasopismach, takich jak *Journal of Machine Learning Research*, *Machine Learning Journal* i *Data Mining and Knowledge Discovery*.

4.2 Lista publikacji

1. Dembczyński, K., Cheng, W., i Hüllermeier, E. (2010a). Bayes optimal multilabel classification via probabilistic classifier chains. W Fürnkranz, J. i Joachims, T., redaktorzy, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, strony 279-286. Omnipress
2. Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2010e). Regret analysis for performance metrics in multi-label classification: The case of Hamming and subset zero-one loss. W Balcázar, J. L., Bonchi, F., Gionis, A., i Sebag, M., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 6321 *Lecture Notes in Computer Science*, strony 280-295. Springer-Verlag
3. Dembczyński, K., Waegeman, W., i Hüllermeier, E. (2012d). An analysis of chaining in multi-label classification. W *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, wolumen 242 *Frontiers in Artificial Intelligence and Applications*, strony 294-299. IOS Press

4. Dembczyński, K., Kotłowski, W., i Hüllermeier, E. (2012b). Consistent multilabel ranking through univariate losses. W Langford, J. i Pineau, J., redaktorzy, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, strony 1319-1326. Omnipress

5. Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2012c). On loss minimization and label dependence in multi-label classification. *Machine Learning*, 88:5-45

6. Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2011). An exact algorithm for F-measure maximization. W Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., i Weinberger, K. Q., redaktorzy, *Advances in Neural Information Processing Systems 24*, strony 1404-1412. Curran Associates, Inc

7. Dembczyński, K., Jachnik, A., Kotłowski, W., Waegeman, W., i Hüllermeier, E. (2013a). Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. W Dasgupta, S. i McAllester, D., redaktorzy, *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, wolumen 28 *Proceedings of Machine Learning Research*, strony 1130-1138. PMLR

8. Waegeman, W., Dembczyński, K., Jachnik, A., Cheng, W., i Hüllermeier, E. (2014). On the Bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research*, 15(1):3333-3388

9. Busa-Fekete, R., Szörényi, B., Dembczyński, K., i Hüllermeier, E. (2015). Online F-measure optimization. W Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., i Garnett, R., redaktorzy, *Advances in Neural Information Processing Systems 28*, strony 595-603. Curran Associates, Inc

10. Jasinska, K., Dembczyński, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., i Hüllermeier, E. (2016). Extreme F-measure maximization using sparse probability estimates. W Balcan, M. F. i Weinberger, K. Q., redaktorzy, *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, wolumen 48 *Proceedings of Machine Learning Research*, strony 1435-1444. PMLR

11. Dembczyński, K., Kotłowski, W., Waegeman, W., Busa-Fekete, R., i Hüllermeier, E. (2016). Consistency of probabilistic classifier trees. W Frasconi, P., Landwehr, N., Manco, G., i Vreeken, J., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 9852 *Lecture Notes in Computer Science*, strony 511-526. Springer

12. Stock, M., Dembczyński, K., Baets, B. D., i Waegeman, W. (2016). Exact and efficient top-k inference for multi-target prediction by querying separable linear relational models. *Data Mining and Knowledge Discovery*, 30(5):1370-1394

4.3 Omówienie celu naukowego ww. prac i osiągniętych wyników wraz z omówieniem ich ewentualnego wykorzystania

4.3.1 Klasyfikacja wieloetykietowa

Klasyfikacja wieloetykietowa jest problemem uczenia maszynowego, w którym wiele etykiet może zostać przypisanych do pojedynczego przykładu. Jest ona naturalnym rozszerzeniem klasyfikacji binarnej oraz wieloklasowej. Problemy wieloetykietowe są powszechnie spotykane w rzeczywistych zastosowaniach. Na przykład film może zostać jednocześnie opisany jako sensacyjny, kryminalny oraz dreszczowiec. Podobnie artykuł prasowy może zostać oznaczony zarazem jako ekonomiczny oraz dotyczący polityki. Natomiast w zastosowaniach biologicznych, dany gen może zostać związany z wieloma klasami funkcjonalnymi, takimi jak metabolizm, transkrypcja czy synteza białek.

Wiele zaproponowanych metod klasyfikacji wieloetykietowej wykorzystuje, w ten czy w inny sposób, zależności pomiędzy etykietami. W porównaniu z prostym podejściem tzw. binarnej stosowności (ang. *binary relevance*), polegającym na uczeniu niezależnego klasyfikatora dla każdej etykiety, jakakolwiek poprawa wyników była zwykle wyjaśniana poprzez fakt, że podejście to ignoruje zależności pomiędzy etykietami. Nie kwestionując poprawności tych badań trzeba przyznać, że to proste wyjaśnienie ukrywa

wiele subtelnych szczegółów i nie prowadzi do znalezienia prawdziwych mechanizmów oraz powodów stojących u podstaw poprawy wyników. Ponadto w problemach wieloetykietowych ze względu na złożoność przestrzeni etykiet możliwe jest zdefiniowanie różnorodnych miar trafności predykcji (lub *funkcji straty* w języku teorii uczenia się). Na przestrzeni lat zaproponowano wiele takich miar, np. *błąd Hamminga*, *wieloetykietowy błąd zerojedynkowy* (ang. *subset 0/1 loss*), *błąd rangowy* (ang. *rank loss*) czy *miarę F*. Pomimo faktu, że powyższe miary mają zupełnie różny charakter, rzadko był pokazywany konkretny związek pomiędzy stosowanym algorytmem uczenia a minimalizowaną funkcją straty, co mylnie sugerowało, że jedna i ta sama metoda może być optymalna dla wielu miar.

W cyklu powyżej wymienionych publikacji udało nam się opracować teoretyczne ramy pozwalające na badanie problemu klasyfikacji wieloetykietowej. Wykorzystując statystyczne spojrzenie na problem wykazaliśmy związek pomiędzy minimalizacją funkcji straty a zależnościami pomiędzy etykietami. Ponadto zaproponowane przez nas algorytmy dla różnych funkcji straty zostały dogłębnie przebadane teoretycznie. Ostatnie nasze wyniki dotyczą problemu tzw. klasyfikacji ekstremalnej, który charakteryzuje się bardzo duża liczbą etykiet (sięgającą nawet milionów). Poniżej opisane są główne osiągnięcia. Odwołania do powyższych prac zostały oznaczone kolorem niebieskim.

4.3.2 Funkcje straty oraz zależności pomiędzy etykietami

Niech $\mathcal{X}$ oznacza przestrzeń przykładów, a $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ skończony zbiór etykiet. Załóżmy, że przykład $\boldsymbol{x} \in \mathcal{X}$ jest powiązany z podzbiorem etykiet $\mathcal{L}_{\boldsymbol{x}} \in 2^{\mathcal{L}}$. Podzbiór ten jest często nazywany zbiorem relewantnych (pozytywnych) etykiet, natomiast jego dopełnienie $\mathcal{L}\backslash\mathcal{L}_{\boldsymbol{x}}$ zbiorem nierelewantnych (negatywnych) etykiet dla $\boldsymbol{x}$. Zbiór $\mathcal{L}_{\boldsymbol{x}}$ jest najczęściej reprezentowany poprzez wektor binarny $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$, taki że $y_i = 1 \Leftrightarrow \lambda_i \in \mathcal{L}_{\boldsymbol{x}}$. Poprzez $\mathcal{Y} = \{0, 1\}^m$ oznaczamy zbiór wszystkich możliwych binarnych wektorów. Ponadto zakładamy, że obserwacje $(\boldsymbol{x}, \boldsymbol{y})$ są generowane niezależnie z tego samego rozkładu $P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})$ zdefiniowanego na przestrzeni $\mathcal{X} \times \mathcal{Y}$.

Problem klasyfikacji wieloetykietowej może zostać zdefiniowany jako poszukiwanie *klasyfikatora* $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x}))$, który minimalizuje *oczekiwaną stratę* (lub inaczej *ryzyko*):

$$L_\ell(\boldsymbol{h}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P(\boldsymbol{x},\boldsymbol{y})}[\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})]\,,$$

gdzie $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$ jest (*zadaniową*) *funkcją straty*. Optymalny klasyfikator, tzw. *klasyfikator bayesowski*, dla danej funkcji straty ℓ jest rozwiązaniem poniższego problemu:

$$\boldsymbol{h}^*_\ell = \arg\min_{\boldsymbol{h}} L_\ell(\boldsymbol{h})\,.$$

Natomiast *żal* klasyfikatora $\boldsymbol{h}$ ze względu na ℓ jest zdefiniowany jako:

$$\mathrm{Reg}_\ell(\boldsymbol{h}) = L_\ell(\boldsymbol{h}) - L_\ell(\boldsymbol{h}^*_\ell) = L_\ell(\boldsymbol{h}) - L^*_\ell\,.$$

Żal określa liczbowo suboptymalność $\boldsymbol{h}$ w porównaniu do klasyfikatora optymalnego $\boldsymbol{h}^*$. Celem klasyfikacji wieloetykietowej jest zatem znalezienie takiego klasyfikatora $\boldsymbol{h}$, którego żal jest jak najmniejszy, w idealnym przypadku równy zero.

Klasyfikator jest uczony na podstawie przykładów uczących $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}^n_1$. Możliwe są dwa podejścia. W pierwszym podejściu poszukiwane jest dobre przybliżenie $\boldsymbol{h}^*$. W drugim podejściu najpierw konstruowany jest model przybliżający rozkład $P(\boldsymbol{y} \mid \boldsymbol{x})$. Następnie podczas klasyfikacji model ten jest wykorzystywany przez procedurę wnioskowania w celu znalezienia najlepszego $\boldsymbol{y}$ dla rozważanej funkcji straty biorąc pod uwagę wyestymowany rozkład warunkowy $P(\boldsymbol{y} \mid \boldsymbol{x})$ dla danego $\boldsymbol{x}$.

Głównymi wyzwaniami klasyfikacji wieloetykietowej jest prawidłowe zamodelowanie zależności pomiędzy etykietami oraz minimalizacja złożonych funkcji straty $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$. W pracach (Dembczyński i inni, 2010a,e, 2012c) rozróżniliśmy dwa rodzaje zależności pomiędzy etykietami, *warunkowe* oraz *brzegowe*. Ten pierwszy rodzaj dotyczy zależności warunkowanych dla danego przykładu $\boldsymbol{x}$, natomiast ten drugi nie odnosi się do konkretnych przykładów i ukazuje globalny (brzegowy) charakter zależności pomiędzy etykietami. W zależności od rozważanego zastosowania oraz używanej funkcji straty należy zamodelować

jeden lub drugi rodzaj zależności. Formalnie, zależności warunkowe są zdefiniowane w następujący sposób:

$$P(\boldsymbol{y} \,|\, \boldsymbol{x}) \neq \prod_{i=1}^{m} P(y_i \,|\, \boldsymbol{x}) \,,$$

natomiast zależności brzegowe jako:

$$P(\boldsymbol{y}) \neq \prod_{i=1}^{m} P(y_i) \,.$$

Należy podkreślić, że zależności brzegowe nie implikują zależności warunkowych. Prawdziwe jest także zdanie odwrotne, tzn. zależności warunkowe nie implikują zależności brzegowych. Warto również zauważyć interesujący fakt, że w przypadku zdegenerowanego rozkładu prawdopodobieństwa (dla którego $P(\boldsymbol{y}) = 1$ dla jednego $\boldsymbol{y}$ oraz $P(\boldsymbol{y}) = 0$ dla wszystkich pozostałych kombinacji etykiet) etykiety są niezależne od siebie.

Źródłem zależności brzegowych mogą być podobieństwa pomiędzy prawdziwymi modelami stojącymi za etykietami. Załóżmy, że prawdziwy model dla i-tej etykiety jest zdefiniowany następująco:

$$h_i(\boldsymbol{x}) = g_i(\boldsymbol{x}) + \epsilon_i(\boldsymbol{x}) \,,$$

gdzie $g_i(\boldsymbol{x})$ jest częścią strukturalną, a $\epsilon_i(\boldsymbol{x})$ częścią losową modelu. Jeżeli występują podobieństwa pomiędzy częściami strukturalnymi $g_i(\boldsymbol{x})$, to warto rozważyć łączne uczenie modeli. Zależności warunkowe z kolei są implikowane poprzez zależności pomiędzy częściami losowymi $\epsilon_i(\boldsymbol{x})$ dla danego $\boldsymbol{x}$.

Funkcje straty możemy podzielić na funkcje dekomponowalne oraz niedekomponowalne ze względu na etykiety:

$$\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \sum_{i=1}^{m} \ell(y_i, h_i(\boldsymbol{x})) \,, \quad \ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) \neq \sum_{i=1}^{m} \ell(y_i, h_i(\boldsymbol{x})) \,.$$

Kanonicznymi przykładami tych dwóch typów funkcji straty są odpowiednio błąd (strata) Hamminga i wieloetykietowy błąd zerojedynkowy:

$$\ell_H(\boldsymbol{y}, \boldsymbol{h}) = \frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h_i]\!] \,, \quad \ell_{0/1}(\boldsymbol{y}, \boldsymbol{h}) = [\![\boldsymbol{y} \neq \boldsymbol{h}]\!] \,.$$

Na podstawie rozważań dotyczących obu rodzajów zależności pomiędzy etykietami oraz obu typów funkcji straty, w pracy (Dembczyński i inni, 2012c) zaproponowaliśmy dwa spojrzenia (ang. *views*) na problem klasyfikacji wieloetykietowej. Pierwsze z nich dotyczy *poszczególnych etykiet* i jest blisko związane z uczeniem się wielu zadań (ang. *multi-task learning*) oraz regresją wielowariacyjną (ang. *multivariate regression*). W tym spojrzeniu głównym zadaniem jest poprawa trafności klasyfikacji dla każdej etykiety z osobna przy wykorzystaniu informacji o innych etykietach. Typową funkcją straty dla takiego zadania jest błąd Hamminga. Zależności brzegowe odgrywają tutaj dużo ważniejszą rolę niż zależności warunkowe. Drugie spojrzenie dotyczy *łącznego rozpatrywania etykiet*, w którym minimalizacja funkcji straty nie może być w prosty sposób rozłożona na wiele niezależnych zadań. Typową funkcją straty dla tego spojrzenia jest wieloetykietowy błąd zerojedynkowy. W tym przypadku zależności warunkowe muszą zostać odpowiednio zamodelowane. W celu zobrazowania różnicy pomiędzy tymi dwoma spojrzeniami wykazaliśmy w pracy (Dembczyński i inni, 2010e, 2012c), że klasyfikator minimalizujący błąd Hamminga może otrzymać prawie dowolnie złe wyniki ze względu na wieloetykietowy błąd zerojedynkowy oraz na odwrót, tzn. klasyfikator minimalizujący wieloetykietowy błąd zerojedynkowy może być prawie dowolnie słaby ze względu na błąd Hamminga. Wynik ten przedstawiony jest pokrótce poniżej.

Można łatwo zauważyć, że klasyfikator bayesowski przyjmuje różne formy dla obu funkcji straty. Niech *ryzyko warunkowe* klasyfikatora $\boldsymbol{h}$ w punkcie $\boldsymbol{x}$ będzie zdefiniowane następująco:

$$L_\ell(\boldsymbol{h} \,|\, \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}} \left[\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) \right] = \sum_{\boldsymbol{y} \in \mathcal{Y}} P(\boldsymbol{y} \,|\, \boldsymbol{x}) \ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) \,.$$

Wtedy optymalny klasyfikator dla danego $\boldsymbol{x}$ jest podany jako:

$$\boldsymbol{h}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{h}} L_\ell(\boldsymbol{h} \,|\, \boldsymbol{x})\,.$$

Okazuje się, że minimalizatorem $\boldsymbol{h}^*_H$ ryzyka Hamminga jest wektor, którego elementami są *mody brzegowe*:

$$h^*_i(\boldsymbol{x}) = \arg\max_{y_i \in \{0,1\}} P(y_i \,|\, \boldsymbol{x})\,, \quad i = 1, \ldots, m\,,$$

natomiast dla wieloetykietowego błędu zerojedynkowego jest to *moda łączna*:

$$\boldsymbol{h}^*_{0/1}(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} P(\boldsymbol{y} \,|\, \boldsymbol{x})\,.$$

W ogólności oba minimalizatory są różne, jednak w niektórych sytuacjach ich wartość może być taka sama, tzn. $\boldsymbol{h}^*_H(\boldsymbol{x}) = \boldsymbol{h}^*_{0/1}(\boldsymbol{x})$. Na przykład wtedy, gdy etykiety $y_1, \ldots, y_m$ są niezależne warunkowo, tzn. gdy $P(\boldsymbol{y} \,|\, \boldsymbol{x}) = \prod_{i=1}^m P(y_i \,|\, \boldsymbol{x})$, lub wtedy, gdy łączna moda spełnia warunek $P(\boldsymbol{h}^*_{0/1}(\boldsymbol{x}) \,|\, \boldsymbol{x}) > 0.5$. Ponadto zachodzi następujące ograniczenie dla dowolnego $\boldsymbol{h}$ i rozkładu $P(\boldsymbol{y} \,|\, \boldsymbol{x})$:

$$\frac{1}{m} L_{0/1}(\boldsymbol{h} \,|\, \boldsymbol{x}) \leqslant L_H(\boldsymbol{h} \,|\, \boldsymbol{x}) \leqslant L_{0/1}(\boldsymbol{h} \,|\, \boldsymbol{x})\,.$$

Powyższe wyniki do pewnego stopnia sugerują, że obie funkcje straty mogą być używane jako swoje zastępniki, ponieważ w niektórych sytuacjach otrzymywane jest to samo rozwiązanie optymalne oraz istnieje wzajemne ograniczenie na wartość ryzyka tych funkcji. Jednakże analiza żalu najgorszego przypadku pokazuje, że minimalizacja wieloetykietowego błędu zerojedynkowego może prowadzić do dużego błędu ze względu na stratę Hamminga i odwrotnie, tzn. minimalizacja straty Hamminga może prowadzić do dużej wartości wieloetykietowego błędu zerojedynkowego. Przypomnijmy, że żal mierzy o ile jest gorszy klasyfikator $\boldsymbol{h}$ w porównaniu z optymalnym klasyfikatorem dla danej funkcji straty. W celu uproszczenia analizy rozważmy *żal warunkowy*:

$$\mathrm{Reg}_\ell(\boldsymbol{h} \,|\, \boldsymbol{x}) = L_\ell(\boldsymbol{h} \,|\, \boldsymbol{x}) - L_\ell(\boldsymbol{h}^*_\ell \,|\, \boldsymbol{x})\,.$$

Poniższa analiza dotyczy żalu pomiędzy klasyfikatorem bayesowskim $\boldsymbol{h}^*_H$ dla straty Hamminga a klasyfikatorem bayesowskim $\boldsymbol{h}^*_{0/1}$ dla wieloetykietowego błędu zerojedynkowego ze względu na obie funkcje straty. Jest to specyficzna analiza, która pokazuje relację pomiędzy dwoma różnymi funkcjami straty.

Poniższe wyniki zostały otrzymany w pracach (Dembczyński i inni, 2010e, 2012c). Pierwsze twierdzenie dotyczy wieloetykietowego błędu zerojedynkowego.

Twierdzenie 1. *Zachodzi następujące ograniczenie górne:*

$$\mathrm{Reg}_{0/1}(\boldsymbol{h}^*_H \,|\, \boldsymbol{x}) = L_{0/1}(\boldsymbol{h}^*_H \,|\, \boldsymbol{x}) - L_{0/1}(\boldsymbol{h}^*_{0/1} \,|\, \boldsymbol{x}) < 0.5\,.$$

Ograniczenie to jest ścisłe, tzn.

$$\sup_P \Big(\mathrm{Reg}_{0/1}(\boldsymbol{h}^*_H \,|\, \boldsymbol{x}) \Big) = 0.5\,,$$

gdzie supremum jest wzięte po wszystkich rozkładach na $\mathcal{Y}$.

Drugie twierdzenie pokazuje podobny wynik dla straty Hamminga.

Twierdzenie 2. *Zachodzi następujące ograniczenie górne dla $m > 3$:*

$$\mathrm{Reg}_H(\boldsymbol{h}^*_{0/1} \,|\, \boldsymbol{x}) = L_H(\boldsymbol{h}^*_{0/1} \,|\, \boldsymbol{x}) - L_H(\boldsymbol{h}^*_H \,|\, \boldsymbol{x}) < \frac{m-2}{m+2}\,.$$

Ograniczenie to jest ścisłe, tzn.

$$\sup_P \Big(\mathrm{Reg}_H(\boldsymbol{h}^*_{0/1} \,|\, \boldsymbol{x}) \Big) = \frac{m-2}{m+2}\,,$$

gdzie supremum jest wzięte po wszystkich rozkładach na $\mathcal{Y}$.

Obie funkcje straty są powszechnie używane w zastosowaniach praktycznych. Jednak warto mieć na uwadze ich ograniczenia. Strata Hamminga nadaje się dobrze do problemów z niezbyt dużą liczbą dobrze zrównoważonych etykiet. Nie powinna być jednak używana w przypadku bardzo dużej liczby etykiet, charakteryzujących się rozkładem o długim ogonie (ang. *long tail distribution*). W takiej sytuacji strata Hamminga będzie przyjmować wartość bliską zeru i posiadać te same wady jak błąd zerojedynkowy w silnie niezrównoważonych problemach binarnych. Dobrym przykładem zastosowania straty Hamminga jest problem predykcji funkcji genów. Wieloetykietowy błąd zerojedynkowy jest bardzo restrykcyjny, jednak może być stosowany w problemach z niewielką liczbą etykiet i niskim poziomem szumu. Tego rodzaju błąd jest często wykorzystywany we wnioskowaniu probabilistycznym (np. wnioskowanie *maximum a posteriori* minimalizuje właśnie ten błąd). Typowym przykładem zastosowania wieloetykietowego błędu zerojedynkowego jest predykcja odpowiedniej kombinacji leków w terapii medycznej.

4.3.3 Algorytmy uczące dla klasyfikacji wieloetykietowej

Uczenie i wnioskowanie przy użyciu wieloetykietowych funkcji straty jest w ogólności trudnym problemem optymalizacyjnym. Funkcje straty, takie jak błąd Hamminga czy wieloetykietowy błąd zerojedynkowy, często nazywane *błędem zadaniowym*, zazwyczaj nie są ani wypukłe ani różniczkowalne. Istnieją jednak dwa podejścia, których celem jest ułatwienie zadania optymalizacji. Są to *redukcja* oraz *minimalizacja zastępczych funkcji straty*. Redukcja polega na przeformułowaniu oryginalnego problemu do prostszych problemów, dla których istnieje efektywne rozwiązanie algorytmiczne. Minimalizacja zastępczych funkcji straty polega na zastąpieniu błędu zadaniowego poprzez stratę/błąd zastępczy, który jest łatwiejszy w optymalizacji. Zazwyczaj zastępcze funkcje straty są wypukłymi i różniczkowalnymi funkcjami, które ograniczają błąd zadaniowy od góry. Przykładami takich funkcji w klasyfikacji binarnej jest strata wykładnicza, logistyczna lub zawiasowa.

Oba podejścia mogą być badane w kontekście zgodności statystycznej, która dotyczy wydajności predykcyjnej algorytmów w przypadku nieskończenie dużego zbioru uczącego (Bartlett i inni, 2006; Tewari i Bartlett, 2007; McAllester, 2009; Gao i Zhou, 2013). Zastępczą funkcję $\tilde{\ell}$ nazywamy *zgodną* (lub *skalibrowaną*) z błędem zadaniowym ℓ, jeżeli zachodzi:

$$\mathrm{Reg}_{\tilde{\ell}}(\boldsymbol{h}) \to 0 \Rightarrow \mathrm{Reg}_{\ell}(\boldsymbol{h}) \to 0\,.$$

Powyższa definicja dotyczy obu podejść, tzn. redukcji oraz minimalizacji zastępczych funkcji straty. W pierwszym przypadku, $\tilde{\ell}$ odpowiada funkcji minimalizowanej w zredukowanym problemie, natomiast w drugim przypadku, $\tilde{\ell}$ jest zastępczą funkcją straty.

Poniżej przedstawione są dwa kanoniczne algorytmy dla problemów klasyfikacji wieloetykietowej, które oparte są na podejściu redukcji. *Binarna stosowność* (ang. *binary relevance*) dekomponuje problem wieloetykietowy do m problemów binarnych, jednego dla każdej etykiety. Algorytm ten wyraźnie upraszcza problem, ponieważ ignorowane są w nim wszelkie zależności pomiędzy etykietami. Powstaje jednak naturalne pytanie, czy istnieje funkcja straty dla której jest on właściwym rozwiązaniem. Drugi algorytm, nazywany *etykietowym zbiorem potęgowym* (ang. *label powerset*), traktuje każdą kombinację etykiet (tzn., wektor $\boldsymbol{y}$) jako nową meta-klasę w problemie wieloklasowym. Dowolny algorytm wieloklasowy może być więc użyty po takim przekształceniu problemu. Niestety liczba tak otrzymanych klas może być bardzo duża (w ogólności jest ona równa 2^m, jednak w praktyce jest ograniczona przez liczbę przykładów uczących). Algorytm ten w odróżnieniu od binarnej stosowności bierze pod uwagę zależności pomiędzy etykietami, jednak ignorowana jest w nim wewnętrzna struktura wektora etykiet $\boldsymbol{y}$.

Na podstawie powyżej przedstawionych wyników teoretycznych dotyczących straty Hamminga i wieloetykietowego błędu zerojedynkowego, nie jest trudno przeanalizować te dwa algorytmy redukcji (Dembczyński i inni, 2010e, 2012c). Łatwo zauważyć, że algorytm binarnej stosowności jest zgodny dla straty Hamminga, bez żadnych dodatkowych założeń dotyczących zależności pomiędzy etykietami. Zauważmy, że jeżeli nie byłoby to prawdą, to nie bylibyśmy w stanie rozwiązań optymalnie żadnego problemu klasyfikacji binarnej (dla etykiety rozważanej w danym problemie istnieją przecież inne etykiety silnie z nią związane). Dla innych funkcji strat konieczne jest zazwyczaj przyjęcie dodatkowych założeń. Na przykład, dla wieloetykietowego błędu zerojedynkowego należy założyć niezależność etykiet lub wysokie

prawdopodobieństwo łącznej mody (> 0.5). Warto zauważyć, że uczenie i wnioskowanie w tym algorytmie jest liniowe z liczbą etykiet.

Algorytm etykietowego zbioru potęgowego jest z kolei zgodny dla wieloetykietowego błędu zerojedynkowego. W jego podstawowej wersji jest on niezgodny dla straty Hamminga. Jednakże przy wykorzystaniu probabilistycznego klasyfikatora wieloklasowego, algorytm ten estymuje łączne prawdopodobieństwo warunkowego dla danego $\boldsymbol{x}$. Dlatego też wnioskowanie dla dowolnego błędu zadaniowego jest możliwe. Podobnie poprzez redukcję do problemu klasyfikacji wieloklasowej z kosztami, algorytm ten może być użyty prawie z każdym błędem zadaniowym. W obydwu przypadkach złożoność obliczeniowa jest jednak wykładnicza z liczbą etykiet.

4.3.4 Probabilistyczne łańcuchy klasyfikatorów

Przy redukcji etykietowego zbioru potęgowego można wykorzystać jeden ze standardowych algorytmów klasyfikacji wieloklasowej, takich jak k-najbliżsi sąsiedzi, drzewa decyzyjne, regresja logistyczna, czy metoda wektorów nośnych. Alternatywnie można skorzystać z dalszej redukcji przekształcając problem wieloklasowy do sekwencji problemów binarnych, używając takich podejść jak 1-przeciw-wszystkim, 1-przeciw-1, ważone wszystkie pary (ang. *weighted all-pairs*) (Beygelzimer i inni, 2008), czy skierowane grafy acykliczne (Platt i inni, 2000). Pozostaje jeszcze jedna możliwość, w której wykorzystywana jest bezpośrednia redukcja do klasyfikacji binarnej, eliminująca problem ignorowania wewnętrznej struktury meta-klas. Można tego dokonać poprzez potraktowanie $\boldsymbol{x}$ oraz $\boldsymbol{y}$ jako cech oraz dodanie nowej zmiennej wyjściowej, która wskazuje czy dany wektor $\boldsymbol{y}$ jest tym prawdziwym dla $\boldsymbol{x}$:

$$(\boldsymbol{x}, \boldsymbol{y}) \longrightarrow \{(\boldsymbol{x}, \boldsymbol{y}, 1)\} \cup \{(\boldsymbol{x}, \boldsymbol{y}', 0) : \forall \boldsymbol{y}' \neq \boldsymbol{y}\}\,.$$

Model w takim podejściu może zostać zdefiniowany poprzez funkcję $f(\boldsymbol{x}, \boldsymbol{y})$. Możliwe są jej różne formy, na przykład:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}, y_i) + \sum_{y_k, y_l} f_{k,l}(y_k, y_l)\,, \tag{1}$$

gdzie drugie wyrażenie dotyczy związków pomiędzy parami etykiet. Predykcja może być wtedy obliczona zgodnie z:

$$\boldsymbol{h}(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})\,.$$

Naiwna implementacja powyższego podejścia jest na pewno nieskalowalna, ponieważ dla każdego przykładu uczącego należy wygenerować wykładniczo wiele przykładów negatywnych (czyli takich, dla których zmienna wyjściowa będzie równa 0). Istnieją jednak dwa podejścia, które w inteligenty sposób rozwiązują ten problem. Pierwsze z nich to losowe pola warunkowe (ang. *conditional random fields*) zaproponowane w Lafferty i inni (2001), które są uogólnieniem regresji logistycznej. Drugie z nich to strukturalna metoda wektorów nośnych (Tsochantaridis i inni, 2005). Niestety, działają one efektywnie tylko dla ograniczonej klasy funkcji $f(\boldsymbol{x}, \boldsymbol{y})$, w której zależności pomiędzy etykietami mogą być zamodelowane poprzez łańcuch linowy lub drzewo o ograniczonej szerokości.

W (Dembczyński i inni, 2010a) udało nam się zaproponować inne podejście. *Probabilistyczne łańcuchy klasyfikatorów* (ang. *probabilistic classifier chains*, PCC) są efektywną metodą redukcji, która uogólnia metodę łańcuchów klasyfikatorów zaproponowaną w (Read i inni, 2009, 2011). PCC estymują łączne prawdopodobieństwo warunkowe $P(\boldsymbol{y} \,|\, \boldsymbol{x})$ poprzez zastosowanie *reguły łańcuchowej*:

$$P(\boldsymbol{y} \,|\, \boldsymbol{x}) = \prod_{i=1}^{m} P(y_i \,|\, y_{i-1}, \ldots, y_1, \boldsymbol{x}) = \prod_{i=1}^{m} P(y_i \,|\, \boldsymbol{y}^{i-1}, \boldsymbol{x})\,,$$

gdzie $\boldsymbol{y}^{i-1} = (y_1, \ldots, y_{i-1})$. Uczenie PCC polega na wytrenowaniu klasyfikatorów probabilistycznych (np., regresji logistycznej) w celu estymacji $P(y_i \,|\, \boldsymbol{y}^{i-1}, \boldsymbol{x})$, niezależnie dla każdego $i = 1, \ldots, m$. Niech

$Q(y_i \mid \boldsymbol{y}^{i-1}, \boldsymbol{x})$ oznacza otrzymane estymaty. Ostateczny model jest wtedy wyrażony poprzez:

$$Q(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{m} Q(y_i \mid \boldsymbol{y}^{i-1}, \boldsymbol{x}) \,.$$

Z teoretycznego punktu widzenia kolejność etykiet w wektorze $\boldsymbol{y}$ nie ma znaczenia. Jednak w praktyce może ona istotnie wpłynąć na trafność nauczonych modeli, ponieważ uczenie przebiega na ograniczonej klasie modeli i na skończonym zbiorze przykładów. Warto również zauważyć, że jeżeli do estymacji $Q(y_i \mid \boldsymbol{y}^{i-1}, \boldsymbol{x})$ zostaną wykorzystane modele liniowe, to ostateczny klasyfikator będzie miał strukturę podobną do funkcji (1). Jednak jego uczenie przebiega w sposób sekwencyjny, a nie bezpośrednio tak jak w losowych polach warunkowych i w strukturalnej metodzie wektorów nośnych.

PCC obliczają w efektywny sposób estymatę prawdopodobieństwa warunkowego dla dowolnego wektora etykiet $\boldsymbol{y}$. W tym celu należy obliczyć $Q(\boldsymbol{y} \mid \boldsymbol{x})$ dla danego $\boldsymbol{y}$, co wymaga odpytania m klasyfikatorów. Jednakże dużo trudniejszym zadaniem jest obliczenie optymalnej (ze względu na rozkład Q) decyzji $\hat{\boldsymbol{y}}^*$ dla danej funkcji straty. W tym celu należy zastosować odpowiednią metodę wnioskowania. Zauważmy, że ostateczny model odpowiada drzewu z binarnymi decyzjami w każdym jego węźle oraz z liśćmi wskazującymi kombinacje etykiet $\boldsymbol{y}$. Każdy węzeł drzewa może zostać jednoznacznie określony poprzez $\boldsymbol{y}^i = (y_1, \ldots, y_i)$. Przy takiej notacji $\boldsymbol{y}^0$ oznacza korzeń drzewa. Najprostszą metodą wnioskowania jest przeszukiwanie zachłanne, które przechodzi tylko jedną ścieżkę od korzenia do liścia drzewa wybierając takie y_i, dla którego $Q(y_i \mid \boldsymbol{y}^{i-1}, \boldsymbol{x})$ jest większe. Przeszukiwanie zachłanne jest szybkie ($O(m)$) i nie wymaga użycia klasyfikatorów probabilistycznych. Jednak jego predykcja nie odpowiada w ogólności ani łącznej ani brzegowej modzie.

W celu znalezienia ścieżki (tzn. kombinacji etykiet $\boldsymbol{y}$) o największym (wyestymowanym) prawdopodobieństwie warunkowym możemy wykorzystać bardziej zaawansowane metody przeszukiwania, takie jak przeszukiwanie wiązkowe (ang. *beam search*) (Kumar i inni, 2013) lub przeszukiwanie z jednolitym kosztem (*uniform-cost search*) (Dembczyński i inni, 2012d). Przy wykorzystaniu tych technik, PCC optymalizują wieloetykietowy błąd zerojedynkowy (dla którego optymalną decyzją jest łączna moda). W pracy (Dembczyński i inni, 2012d) udało nam się zaproponować efektywny algorytm, nazwany ϵ-przybliżonym wnioskowaniem, który jest wariantem przeszukiwania z jednolitym kosztem z punktem odcięcia. Algorytm ten zawsze znajdzie łączną modę rozkładu Q, jeżeli jej prawdopodobieństwo jest większe lub równe ϵ. Ponadto udało nam się udowodnić następujące twierdzenie (poniżej przedstawiona jest jego lekko zmodyfikowana wersja w porównaniu z oryginalną pracą)

Twierdzenie 3. *Niech $1 \leqslant c \leqslant m$. Algorytm ϵ-przybliżonego wnioskowania dla $\epsilon = 2^{-c}$ potrzebuje co najwyżej $O(m\epsilon^{-1})$ iteracji, aby znaleźć predykcję $\boldsymbol{h}_\epsilon(\boldsymbol{x}) = \hat{\boldsymbol{y}}_\epsilon$ dla której*

$$Q(\hat{\boldsymbol{y}}^* \mid \boldsymbol{x}) - Q(\hat{\boldsymbol{y}}_\epsilon \mid \boldsymbol{x}) \leqslant \epsilon - 2^{-m} \,,$$

gdzie $\hat{\boldsymbol{y}}^* = \arg\max_{\boldsymbol{y}} Q(\boldsymbol{y} \mid \boldsymbol{x})$.

Bazując na powyższym wyniku można także pokazać, że rozwiązanie optymalne jest znajdowane w czasie liniowym od $1/p_{\max}$, gdzie $p_{\max}$ jest prawdopodobieństwem łącznej mody. Dla problemów z małym szumem (czyli z dużymi wartościami $p_{\max}$), algorytm ten będzie działał bardzo szybko. Zauważmy, że przeszukiwanie zachłanne, które odpowiada powyższemu algorytmowi z $\epsilon = 0.5$, charakteryzuje się słabymi gwarancjami:

$$Q(\hat{\boldsymbol{y}}^* \mid \boldsymbol{x}) - Q(\hat{\boldsymbol{y}}_{greedy} \mid \boldsymbol{x}) \leqslant 0.5 - 2^{-m} \,.$$

W jednej z ostatnich prac (Dembczyński i inni, 2016), rozszerzyliśmy powyższe wyniki. Wykazaliśmy, że koncepcja stojąca za algorytmem PCC może być również zastosowana do problemów klasyfikacji wieloklasowej, jeżeli etykiety wieloklasowe zostaną zastąpione poprzez binarny kod przedrostkowy. Taki kod może być zawsze przedstawiony jako drzewo binarne. W węzłach wewnętrznych takiego drzewa trenowany jest osobny klasyfikator binarny. To jest jedna z głównych różnic w porównaniu z algorytmem PCC, w którym klasyfikator binarny jest budowany na każdym poziomie drzewa (ponieważ poziom drzewa odpowiada etykiecie). Zaproponowane podejście zostało przez nas nazwane *probabilistycznym drzewem klasyfikatorów*. Ponadto uogólniliśmy wyniki teoretyczne dotyczące algorytmu ϵ-przybliżonego wnioskowania do przeszukiwania A*.

Udało nam się także udowodnić górne ograniczenie żalu dla błędu zerojedynkowego. Ograniczenie to jest wyrażone poprzez błąd indywidualnych klasyfikatorów oraz przybliżenie wynikające z metody przeszukiwania. Nasz główny wynik podany poniżej dotyczy zarówno problemów wieloetykietowych, jak i wieloklasowych. Niech m oznacza liczbę etykiet w pierwszym przypadku oraz długość najdłuższego kodu w drugim przypadku. Przypomnijmy, że $\boldsymbol{y}^i = (y_1, \ldots, y_i)$ jednoznacznie określa węzeł w drzewie, gdzie $\boldsymbol{y}^0$ oznacza korzeń drzewa.

Twierdzenie 4. *Niech $Q(\cdot \,|\, \boldsymbol{y}^{i-1}, \boldsymbol{x})$ będą estymatami prawdopodobieństw warunkowych w każdym wewnętrznym węźle drzewa $\boldsymbol{y}^{i-1}$ oraz niech $\boldsymbol{h}_\epsilon$ będzie takim klasyfikatorem, który dla danego $\boldsymbol{x}$ przewiduje wektor $\hat{\boldsymbol{y}}_\epsilon$ znaleziony przez ϵ-przybliżone wnioskowanie. Wtedy, dla dowolnego rozkładu P,*

$$\mathrm{Reg}_{0/1}(\boldsymbol{h}_\epsilon) \leqslant \sqrt{2\mathrm{Reg}_{\log}(Q)} + \epsilon - 2^{-m}\,,$$

gdzie $\mathrm{Reg}_{\log}(Q) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P}\left[\sum_{i=1}^{m} \mathrm{Reg}_{\log}(Q(\cdot \,|\, \boldsymbol{y}^{i-1}, \boldsymbol{x}))\right]$ jest oczekiwaną sumą żalu logistycznego wewnętrznych klasyfikatorów binarnych na ścieżce od korzenia do liścia.

Co ciekawe, koncepcja podobna do probabilistycznych drzew klasyfikatorów została niezależnie zaproponowana w wielu różnych dziedzinach badawczych. W sieciach głębokich, takie podejście jest znane pod nazwą hierarchicznego softmaksu (Morin i Bengio, 2005), w statystyce jako zagnieżdżone dychotomie (Fox, 1997), natomiast w wieloklasowej regresji jako drzewa estymujące prawdopodobieństwo warunkowe (Beygelzimer i inni, 2009). W rozpoznawaniu wzorców podobne podejście było rozważane jako klasyfikatory wieloetapowe (ang. *multi-stage*) (Kurzynski, 1988). Powyższe wyniki teoretyczne dotyczą w ogólności wszystkich wyżej wymieniowych podejść. Innymi słowy, udało nam się zunifikować podobne podejścia w celu wykazania ich własności teoretycznych.

Dotychczasowa dyskusja dotyczyła głównie (wieloetykietowego) błędu zerojedynkowego. W ogólności PCC estymują łączne prawdopodobieństwo warunkowe w punkcie $\boldsymbol{x}$. Dlatego też możliwe jest zastosowanie wnioskowanie dla dowolnej funkcji straty. W tym celu należy przejść całe drzewo w celu otrzymania pełnego rozkładu $Q(\boldsymbol{y} \,|\, \boldsymbol{x})$. Następnie należy zastosować odpowiednie wnioskowanie dla danej funkcji straty ℓ:

$$\hat{y}^* = \arg\max_{\boldsymbol{h}\in\mathcal{Y}} \sum_{\boldsymbol{y}\in\mathcal{Y}} Q(\boldsymbol{y} \,|\, \boldsymbol{x})\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))\,.$$

Niestety takie podejście jest bardzo kosztowne, ponieważ pełen rozkład dotyczy wszystkich 2^m kombinacji etykiet. W celu obniżenia kosztów obliczeniowych, można wykorzystać próbkowanie ancestralne (ang. *ancestral sampling*). Wnioskowanie przebiega wtedy na otrzymanej próbce, czasami nazywanej rozkładem empirycznym. Dla straty Hamminga, na przykład, należy wyestymować z tej próbki brzegowe prawdopodobieństwa warunkowe. Niestety dla pewnych funkcji straty obliczenie optymalnej decyzji na podstawie próbki może być ciągle bardzo kosztowne.

Warto podkreślić, że zaproponowany przez nas algorytm PCC jest dobrze rozpoznawany w środowisku naukowym i traktowany jako wzorcowe podejście do klasyfikacji wieloetykietowej z błędem zerojedynkowym, będące silną konkurencją dla takich algorytmów jak strukturalna metoda wektorów nośnych. Można także wykazać podobieństwo PCC do modeli Markowa o maksymalnej entropii (ang. *maximum entropy Markov models*) (McCallum i inni, 2000) zaproponowanych dla problemu etykietowania sekwencji oraz do algorytmów opierających się na nowym paradygmacie uczenia się przeszukiwania (ang. *learn to search*) (Daumé III i inni, 2009; Doppa i inni, 2014).

4.3.5 Algorytmy dla innych błędów zadaniowych

W naszych badaniach dogłębnie przeanalizowaliśmy także dwa inne błędy zadaniowe: miarę F_β oraz błąd rangowy. W obydwóch przypadkach udało nam się otrzymać znaczące wyniki.

Miara F_β jest powszechnie używana w wyszukiwaniu informacji, tagowaniu dokumentów oraz w przetwarzaniu języka naturalnego. Funkcja straty oparta na tej mierze jest zdefiniowana następująco:

$$\ell_{F_\beta}(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) \quad = \quad 1 - F_\beta(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = 1 - \frac{(1+\beta^2)\sum_{i=1}^{m} y_i h_i(\boldsymbol{x})}{\beta^2 \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} h_i(\boldsymbol{x})}\,.$$

Miara F_β zapewnia lepsze niż strata Hamminga zrównoważenie relewantnych i nierelewantnych etykiet. Jej optymalizacja jest jednak dużym wyzwaniem. Wcześniejsze algorytmy były w stanie tylko przybliżyć optymalne rozwiązanie lub bazowały na dodatkowych założeniach dotyczących rozkładu prawdopodobieństwa etykiet (np. zakłady ich niezależność). W (Dembczyński i inni, 2011) znaleźliśmy rozwiązanie, które jest zarazem statystycznie zgodne oraz obliczeniowo efektywne. Bazuje ono na metodzie podstawień (ang. *plug-in estimate*), w której odpowiednie estymaty są podstawiane do wyrażenia na klasyfikator bayesowski dla danej funkcji straty. Poniżej przedstawione jest krótkie podsumowanie tej metody.

Rozważany przez nas problem może zostać sformułowany następująco:

$$\boldsymbol{h}^*(\boldsymbol{x}) \quad = \quad \arg\min_{\boldsymbol{h}\in\mathcal{Y}} \mathbb{E}\left[\ell_{F_\beta}(\boldsymbol{Y},\boldsymbol{h}(\boldsymbol{x}))\right] = \arg\max_{\boldsymbol{h}\in\mathcal{Y}} \sum_{\boldsymbol{y}\in\mathcal{Y}} P(\boldsymbol{y}\,|\,\boldsymbol{x}) \frac{(\beta+1)\sum_{i=1}^m y_i h_i(\boldsymbol{x})}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i(\boldsymbol{x})}\,. \tag{2}$$

Niestety nie istnieje rozwiązanie analityczne powyższego problemu optymalizacji. Także rozwiązanie siłowe jest w ogólnym przypadku niewykonalne, gdyż wymagałoby ono sprawdzenia rozwiązania dla 2^m możliwych kombinacji etykiet, dla każdej z nich policzenia sumy po 2^m komponentach w celu otrzymania wartości oczekiwanej, a także wyestymowania 2^m parametrów, czyli pełnego rozkładu warunkowego $P(\boldsymbol{y}\,|\,\boldsymbol{x})$. Okazuje się jednak, że istnieje prosty algorytm, który rozwiązuje powyższy problem w efektywny sposób. W pracy (Dembczyński i inni, 2011) wykazaliśmy, że jedynie m^2+1 parametrów łącznego rozkładu $P(\boldsymbol{y}\,|\,\boldsymbol{x})$ jest koniecznych do otrzymania rozwiązania.

Twierdzenie 5. *Niech* $s_{\boldsymbol{y}} = \sum_{i=1}^m y_i$. *Rozwiązanie problemu (2) może zostać obliczone na podstawie* $P(\boldsymbol{y}=\boldsymbol{0}\,|\,\boldsymbol{x})$ *oraz wartości*

$$p_{is} = P(y_i = 1, s_{\boldsymbol{y}} = s), \quad i,s \in \{1,\ldots,m\}\,,$$

które stanowią macierz P *o rozmiarze* $m \times m$.

Co ciekawe, elementy macierzy P nie dotyczą zależności pomiędzy etykietami, a odnoszą się raczej do wartości brzegowych, które uwzględniają liczbę wspólnie występujących etykiet (bez wskazywania, które są to etykiety). Algorytm przez nas zaproponowany, nazwany jako *ogólny maksymalizator miary* F_β (ang. *general* F_β*-measure maximizer*, GFM) oblicza rozwiązanie optymalne w czasie sześciennym.

Twierdzenie 6. *GFM rozwiązuje problem (2) w czasie* $o(m^3)$ *zakładając, że macierz* P *zawierająca* m^2 *wartości oraz* $P(\boldsymbol{y}=0\,|\,\boldsymbol{x})$ *są podane.*

Powyższy algorytm wnioskowania może zostać wykorzystany razem z algorytmem PCC. W celu obliczenia optymalnej decyzji dla danego $\boldsymbol{x}$ próbkujemy n obserwacji z rozkładu $Q(\boldsymbol{y}\,|\,\boldsymbol{x})$ otrzymanego przez PCC. Na podstawie tej próbki estymujemy wartości macierzy P oraz prawdopodobieństwo $P(\boldsymbol{y}=\boldsymbol{0}\,|\,\boldsymbol{x})$. Następnie wartości te są wykorzystywane przez algorytm GFM. Wyniki eksperymentalne pokazują, że to podejście osiąga wyniki lepsze od konkurencyjnych algorytmów.

Powyższe podejście zostało przez nas dalej rozszerzone w pracy (Dembczyński i inni, 2013a). Zaproponowaliśmy metodę, która zamiast algorytmu PCC wykorzystuje wielomianowe modele regresyjne do obliczenia parametrów wymaganych przez algorytm GFM. Rozważaliśmy również uproszczony wariant algorytmu, który zakłada niezależność etykiet. W takim przypadku wystarczy wyestymować prawdopodobieństwa brzegowe przy użyciu podejścia binarnej stosowności z klasyfikatorem probabilistycznym, np. regresją logistyczną. Wnioskowanie jest wtedy przeprowadzane zgodnie z algorytmem zaproponowanym w (Ye i inni, 2012), którego złożoność jest kwadratowa dla wymiernego β oraz sześcienna w ogólnym przypadku. Ponadto pokazaliśmy eksperymentalnie, że zaproponowane przez nas algorytmy przewyższają podeście oparte na strukturalnej metodzie wektorów nośnych przystosowane do maksymalizacji miary F_β (Petterson i Caetano, 2010, 2011). Przeprowadziliśmy również analizę teoretyczną obu podejść, która wykazała, że podejście bazujące na wielomianowej regresji i algorytmie GFM jest statystycznie zgodne, natomiast strukturalna metoda wektorów nośnych nie jest. Ten wynik podkreśla jeszcze mocniej znaczenie naszego wkładu w rozwój algorytmów klasyfikacji wieloetykietowej.

Wyniki dotyczące maksymalizacji miary F_β zostały przez nas podsumowane w pracy (Waegeman i inni, 2014). Dodatkowo wykazaliśmy w tym artykule, że algorytm GFM wymaga jedynie czasu kwadratowego, jeżeli jego parametrem wejściowym jest macierz Δ, będąca iloczynem macierzy P oraz macierzy W

składającej się z elementów $w_{sk} = (s+k)^{-1}$, $s, k \in \{1, \ldots, m\}$. Przeanalizowaliśmy również związek pomiędzy klasyfikatorem bayesowskim dla miary F_β a klasyfikatorami bayesowskimi dla innych funkcji straty, takich jak strata Hamminga, wieloetykietowy błąd zerojedynkowy oraz współczynnik Jaccarda. Wykazaliśmy jednoznacznie, że żadna z tych funkcji straty nie może zastąpić miary F_β bez wyraźnego pogorszenia trafności predykcji. Ten wynik otrzymaliśmy poprzez analizę żalu podobną do tej, którą użyliśmy do zbadania relacji pomiędzy stratą Hamminga a wieloetykietowym błędem zerojedynkowym. Omówiliśmy także dogłębnie wykorzystanie algorytmu GFM z takimi metodami uczenia maszynowego jak drzewa decyzyjne, k-najbliżsi sąsiedzi, PCC i wielomianowa regresja.

Błąd rangowy mierzy niezgodność pomiędzy parami obserwacji. W przypadku klasyfikacji binarnej jego odpowiednikiem jest pole powierzchni pod krzywą ROC. W klasyfikacji wieloetykietowej jest zazwyczaj używany na poziomie pojedynczego przykładu do zliczania liczby niezgodności pomiędzy parami etykiet. Jest on wtedy zdefiniowany w następujący sposób:

$$\ell_{\mathrm{rnk}}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x})) = w(\boldsymbol{y}) \sum_{(i,j):\, y_i > y_j} \left([\![f_i(\boldsymbol{x}) < f_j(\boldsymbol{x})]\!] + \frac{1}{2} [\![f_i(\boldsymbol{x}) = f_j(\boldsymbol{x})]\!] \right),$$

gdzie $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$ jest wektorem funkcji, które przypisują wartość rzeczywistą do poszczególnych etykiet, a $w(\boldsymbol{y})$ jest funkcją ważącą, która może zostać wykorzystana do normalizacji błędu rangowego do przedziału $[0, 1]$. W takim przypadku funkcja ta jest równa odwrotności iloczynu liczby n_+ relewantnych i nierelewantnych n_- etykiet, czyli $w(\boldsymbol{y}) = (n_+ n_-)^{-1}$.

Co ciekawe, bardzo intuicyjne podejście oparte na zastosowaniu wypukłych funkcji zastępczych zdefiniowanych na parach etykiet o postaci

$$\tilde{\ell}_\phi(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x})) = \sum_{(i,j):\, y_i > y_j} w(\boldsymbol{y}) \phi(f_i(\boldsymbol{x}) - f_j(\boldsymbol{x})),$$

gdzie ϕ jest np. stratą wykładniczą, logistyczną lub zawiasową, jest niezgodne statystycznie (Duchi i inni, 2010; Gao i Zhou, 2013). Zgodny klasyfikator może być jednak otrzymany poprzez minimalizację prostych funkcji straty, co udało nam się wykazać w (Dembczyński i inni, 2012b).

Okazuje się, że nawet w ogólnym przypadku klasyfikator bayesowski dla błędu rangowego można otrzymać poprzez posortowanie etykiet zgodnie z następującymi wielkościami brzegowymi:

$$\Delta_i^1 = \sum_{\boldsymbol{y}:\, y_i = 1} w(\boldsymbol{y}) P(\boldsymbol{y} \,|\, \boldsymbol{x}).$$

Dla $w(\boldsymbol{y}) \equiv 1$, Δ_i^u redukuje się do prawdopodobieństwa brzegowego $P(y_i = u \,|\, \boldsymbol{x})$ (Dembczyński i inni, 2010a). Powyższy wynik sugeruje, że w celu rozwiązania problemu wystarczy wykorzystać podejście binarnej stosowności, a dokładniej jego ważony wariant.

Rozważmy następujące funkcje straty będące ważonym wariantem straty wykładniczej i logistycznej:

$$\begin{aligned} \tilde{\ell}_{\exp}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x})) &= w(\boldsymbol{y}) \sum_{i=1}^{m} e^{-y_i' f_i(\boldsymbol{x})}, \\ \tilde{\ell}_{\log}(\boldsymbol{y}, \boldsymbol{f}(\boldsymbol{x})) &= w(\boldsymbol{y}) \sum_{i=1}^{m} \log\left(1 + e^{-y_i' f_i(\boldsymbol{x})}\right), \end{aligned}$$

gdzie $y' = 2y - 1$. Minimalizacja powyższych funkcji sprowadza się do rozwiązania m niezależnych problemów, jednego dla każdej etykiety. Dowolny algorytm klasyfikacji binarnej wykorzystujący stratę wykładniczą lub logistyczną, na przykład AdaBoost lub regresja logistyczna, może być tutaj wykorzystany, pod warunkiem że pozwala on na uczenie z ważonymi przykładami. Powyższe podejście do minimalizacji błędu rangowego, pomimo swojej prostoty i efektywności, jest także statystycznie zgodne jak wykazuje poniższy wynik oryginalnie zaprezentowany w pracy (Dembczyński i inni, 2012b).

Minimalizatorem powyższych funkcji straty jest:

$$f_i^*(\boldsymbol{x}) = \frac{1}{c}\log\frac{\Delta_i^1}{\Delta_i^0} = \frac{1}{c}\log\frac{\Delta_i^1}{W-\Delta_i^1}\,,$$

gdzie $c = 1$ dla straty logistycznej, $c = 2$ dla straty wykładniczej, oraz W jest oczekiwaną wagą, tzn.:

$$W = \mathbb{E}_{\boldsymbol{y}}[w(\boldsymbol{y}) \mid \boldsymbol{x}] = \sum_{\boldsymbol{y}} w(\boldsymbol{y}) P(\boldsymbol{y} \mid \boldsymbol{x})\,.$$

Zauważmy, że powyższy minimalizator jest ściśle rosnącym przekształceniem Δ_i^1. Na podstawie powyższych wyników, udało nam się następnie udowodnić następujące ograniczenie górne na żal błędu rangowego.

Twierdzenie 7. *Niech* $\mathrm{Reg}_{\exp}(\boldsymbol{f})$ *i* $\mathrm{Reg}_{\log}(\boldsymbol{f})$ *będą odpowiednio żalem dla ważonej straty wykładniczej i logistycznej. Wtedy*

$$\begin{aligned}\mathrm{Reg}_{\mathrm{rnk}}(\boldsymbol{f}) &\leqslant \frac{\sqrt{6}}{4} C \sqrt{\mathrm{Reg}_{\exp}(\boldsymbol{f})}\,,\\ \mathrm{Reg}_{\mathrm{rnk}}(\boldsymbol{f}) &\leqslant \frac{\sqrt{2}}{4} C \sqrt{\mathrm{Reg}_{\log}(\boldsymbol{f})}\,,\end{aligned}$$

gdzie $C \leqslant m\sqrt{m w_{\max}}$*, z* $w_{max} > w(\boldsymbol{y})$ *dla każdego* $\boldsymbol{y}$*.*

4.3.6 Klasyfikacja ekstremalna

Praktyczne problemy uczenia maszynowego charakteryzują się obecnie nie tylko ogromną liczbą przypadków uczących i cech użytych do ich opisu, ale także bardzo dużą liczbą klas (etykiet), do których przykłady są przypisywane. Takie problemy są często nazywane klasyfikacją ekstremalną. Mogą one mieć zarówno charakter wieloklasowy jak i wieloetykietowy. Przykładowym zastosowaniem klasyfikacji ekstremalnej jest etykietowanie zdjęć i filmów (np. w celu ułatwienia wyszukiwania tego typu plików), tagowanie dokumentów tekstowych (np. do automatycznej kategoryzacji artykułów Wikipedii), rekomendacja słów kluczowych w reklamie internetowej lub predykcja następnego słowa w wypowiedzi. Klasyfikacja ekstremalna pozwala również spojrzeć z innej perspektywy na takie problemy jak ranking czy rekomendacja poprzez sformułowanie ich jako problem wieloetykietowy, w którym rangowany element lub rekomendowany produkt jest traktowany jako osobna etykieta.

W celu lepszego zobrazowania problemu klasyfikacji ekstremalnej rozpatrzmy konkretny problem tagowania artykułów Wikipedii. W tym przypadku pojedynczy artykuł jest przykładem/obserwacją, słowa w nim występujące odpowiadają cechom, natomiast kategorie, do których jest on przypisany, etykietom. Zbiór danych stworzony na podstawie bieżącej zawartości Wikipedii będzie charakteryzował się milionami przykładów i cech, ale także ponad milionem etykiet, ponieważ tak dużo kategorii jest obecnie używanych.

Wyzwania, które niosą ze sobą powyższe problemy, otworzyły nowy kierunek badań w obrębie uczenia maszynowego. Łatwo zauważyć, że podejście binarnej stosowności (czy też 1-przeciw-wszystkim, w przypadku klasyfikacji wieloklasowej), które skaluje się liniowo z liczbą etykiet, jest zbyt kosztowe w przypadku problemów z tak dużą liczbą klas. Dlatego też wymagane jest zaproponowanie nowych algorytmów o subliniowej złożoności czasowej i pamięciowej.

Udało nam się zaproponować (Jasinska i inni, 2016) nowy algorytm efektywnie rozwiązujący problem ekstremalnej klasyfikacji wieloetykietowej, nazwany *probabilistycznymi drzewami etykiet* (ang. *probabilistic label trees*, PLT). Może on być łatwo dostosowany do różnych miar trafności predykcji, takich jak strata Hamminga, precyzja na k-tym miejscu, czy makro-uśredniona miara F_β. Bazuje on na tej samej koncepcji co probabilistyczne drzewa klasyfikatorów (oraz pozostałe podobne algorytmy, takie jak np. hierarchiczny softmax), przez co może on zostać uznany za odpowiednie rozszerzenie tego podejścia do estymacji prawdopodobieństw brzegowych etykiet.

PLT wykorzystuje klasyfikatory probabilistyczne w każdym węźle drzewa w taki sposób, aby estymata prawdopodobieństwa etykiety związanej z danym liściem drzewa była wyrażona iloczynem estymat

prawdopodobieństw otrzymanych na ścieżce od korzenia do tego liścia. Predykcja polega zatem na przeszukiwaniu drzewa od korzenia do odpowiednich liści. Jeżeli pośrednia wartość tego iloczynu na ścieżce od korzenia do danego węzła wewnętrznego jest mniejsza od zadanego progu, to poddrzewo zaczynające się w tym węźle nie jest dalej przeszukiwane. Ta strategia odcinania poddrzew z etykietami o zbyt małym prawdopodobieństwie brzegowym prowadzi do bardzo szybkiej procedury klasyfikacji nowych przykładów.

Wzór na prawdopodobieństwo brzegowe $P(y_i = 1 \,|\, \boldsymbol{x})$ może zostać przedstawiony w następujący sposób:

$$P(y_i = 1 \,|\, \boldsymbol{x}) = \prod_{t \in \mathrm{Path}(i)} P(z_t = 1 \,|\, z_{\mathrm{pa}(t)} = 1, \boldsymbol{x}) \,,$$

gdzie $\mathrm{Path}(i)$ jest ścieżką od korzenia do liścia i, z_t jest zmienną wyjściową w węźle t równą 1 wtedy i tylko wtedy, gdy w poddrzewie węzła t jest przynajmniej jedna etykieta pozytywna, a $\mathrm{pa}(t)$ jest rodzicem węzła t. Zakładamy, że w przypadku korzenia $P(z_t = 1 \,|\, z_{\mathrm{pa}(t)} = 1, \boldsymbol{x}) = P(z_t = 1 \,|\, \boldsymbol{x})$. Poprawność wyrażenia na prawdopodobieństwo brzegowe wynika z faktu, że $z_t = 1$ implikuje $z_{\mathrm{pa}(t)} = 1$.

Uczenie PLT może przebiegać w trybie wsadowym lub przyrostowym. Ze względu na warunek użyty w węzłach drzewa (tzn., $z_{\mathrm{pa}(t)} = 1$), dany przykład uczący jest wykorzystywany do uczenia tylko niektórych klasyfikatorów wewnętrznych. Dzięki temu procedura uczenia jest bardzo efektywna. Ponadto poprzez wykorzystanie haszowania cech (ang. *feature hashing*) (Weinberger i inni, 2009) wszystkie klasyfikatory mogą być składowane we wspólnej przestrzeni pamięciowej o kontrolowanym rozmiarze. Udało nam się także wyprowadzić następujące ograniczenia na żal:

$$|P(y_i = 1 \,|\, \boldsymbol{x}) - Q(y_i = 1 \,|\, \boldsymbol{x})| \leqslant \sum_{t \in \mathrm{Path}(i)} \sqrt{\frac{2}{\lambda}} \sqrt{\mathrm{Reg}_\ell(f_t \,|\, \boldsymbol{x})} \,,$$

gdzie Q jest wyestymowanym prawdopodobieństwem brzegowym, ℓ jest silnie prawidłowo złożoną zastępczą funkcją straty (ang. *strongly proper composite surrogate loss*) taką jak np. strata kwadratowa, wykładnicza lub logistyczna, wykorzystaną do uczenia klasyfikatorów wewnętrznych f_t oraz λ jest stałą charakteryzującą daną funkcję straty (Agarwal, 2014).

W celu predykcji k etykiet o największym (wyestymowanym) prawdopodobieństwie brzegowym, a co za tym idzie optymalizacji precyzji na k-tym miejscu, wystarczy wykorzystać kolejkę priorytetową do przeszukiwania drzewa. W eksperymencie obliczeniowym udało nam się pokazać (Jasinska i inni, 2016), że ten wariant PLT uzyskuje wyniki konkurencyjne do algorytmu FastXML, będącego jednym z flagowych podejść rozwiązujących problem klasyfikacji ekstremalnej (Prabhu i Varma, 2014).

PLT może być także wykorzystany do optymalizacji makro-uśrednionej miary F_β. Jest to możliwe, ponieważ procedura przeszukiwania drzewa może być użyta z dowolnym progiem w każdym węźle drzewa. Dzięki temu możemy zdefiniować różne progi odcięcia dla każdej etykiety. Okazuje się, że jest to dokładnie to, co jest potrzebne do optymalizacji makro uśrednionej miary F_β. Należy znaleźć optymalny próg na prawdopodobieństwach warunkowych dla każdej etykiety z osobna. Ponadto, wyznaczenie optymalnego progu może być wykonane efektywnie za pomocą algorytmu przyrostowego zaproponowanego przez nas w pracy (Busa-Fekete i inni, 2015). Zakładając, że przykłady przychodzą w sposób sekwencyjny, zaproponowany algorytm stara się maksymalizować tzw. *przyrostową miarę* F, zdefiniowaną następująco dla przykładu j oraz etykiety i:

$$F_{j,i} = \frac{2 \sum_{t=1}^{j} y_{t,i} \hat{y}_{t,i}}{\sum_{t=1}^{j} y_{t,i} + \sum_{t=1}^{j} \hat{y}_{t,i}} = \frac{2a_{j,i}}{b_{j,i}} \,.$$

Algorytm ustawia próg

$$\tau_{j,i} = \frac{a_{j-1,i}}{b_{j-1,i}}$$

podczas przetwarzania j-tego przykładu i tym samym dokonuje predykcji zgodnie z $\hat{y}_{j,i} = [\![Q(y_i = 1 \,|\, \boldsymbol{x}_j) > \tau_{j,i}]\!]$. Udało nam się udowodnić w (Busa-Fekete i inni, 2015), że tak ustawiany próg zbiega według prawdopodobieństwa do optymalnej wartości, tzn. $\tau_{j,i} \xrightarrow{P} \tau_i^*$. Ponieważ powyższy algorytm bazuje tylko na pozytywnych etykietach (tzn. $y_i = 1$) oraz etykietach z pozytywną predykcją (tzn. $\hat{y}_i = 1$), może on

być łatwo zastosowany razem z algorytmem PLT do klasyfikacji ekstremalnej z makro uśrednioną miarą F_β.

W pracy (Stock i inni, 2016) rozważaliśmy ogólny problem predykcji z wieloma celami (ang. *multi-target prediction*). W ramach tego problemu interesowały nas strategie efektywnej i dokładnej predykcji k najlepszych odpowiedzi, które nie wymagają wykonania obliczeń dla wszystkich celów. Zdefiniowaliśmy klasę *separowalnych relacyjnych modeli linowych* (ang. *separable linear relational models*) i wykazaliśmy, że algorytm progowy zaproponowany w (Fagin i inni, 2003) może być skutecznie użyty z tą klasą modeli.

Rozważmy dwa typy obiektów $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Zaproponowany model dla każdej pary (x, y) oblicza następującą wartość jako predykcję:

$$s(x, y) = \boldsymbol{u}(x)^\top \boldsymbol{t}(y) = \sum_{r=1}^{R} u_r(x) t_r(y) \,.$$

Obiekty x oraz y są reprezentowane odpowiednio poprzez R-wymiarowy model $\boldsymbol{u}$ oraz $\boldsymbol{t}$. Zauważmy, że powyższe sformułowanie jest bardzo ogólne i spotykane w wielu problemach. Na przykład w faktoryzacji macierzy R odpowiada rzędowi macierzy użytej do dekompozycji oryginalnej macierzy. W problemie klasyfikacji wieloetykietowej R odpowiada liczbie cech, a $s(x, y)$ predykcji dla danej etykiety. W problemie wyszukiwania najbliższych sąsiadów x oraz y pochodzą z tej samej dziedziny o wymiarze R.

Udało nam się wykazać, że algorytm progowy jest nadal optymalny dla każdej instancji problemu (ang. *instance optimal*) w przypadku separowalnych relacyjnych modeli linowych. Ten rodzaj optymalności oznacza, że dla każdej instancji problemu nie ma innego dokładnego algorytmu o niższej złożoności czasowej (z dokładnością do stałej). Ponadto wyniki eksperymentalne wskazują, że jest to właściwe podejście do dokładnego wyszukiwania k najlepszych odpowiedzi. Niestety, w przypadku bardzo dużych przestrzeni wyjść konieczne jest wykorzystanie algorytmów przybliżonych.

5 Omówienie pozostałych osiągnięć naukowo-badawczych

5.1 Modele regułowe

Problem indukcji reguł decyzyjnych odgrywa ważną rolę w uczeniu maszynowym. Główną zaletą reguł decyzyjnych jest ich prostota oraz łatwo interpretowana przez człowieka postać. Ponadto pozwalają one na modelowanie złożonych zależności pomiędzy cechami. Kontynuując badania będące główną częścią mojej pracy doktorskiej, zaproponowaliśmy oraz dokładnie przeanalizowaliśmy algorytm indukcji reguł decyzyjnych o nazwie ENDER (Dembczyński i inni, 2010c). Algorytm ten jest dostosowany zarówno do problemu regresji jak i klasyfikacji binarnej. Wykorzystuje do uczenia podejście zwiększania gradientu (ang. *gradient boosting*), które może być traktowane jako uogólnienie sekwencyjnego pokrywania, będącego najbardziej popularnym podejściem wykorzystywanym do uczenia się reguł. W pracy rozważaliśmy różne funkcje straty oraz algorytmy ich minimalizacji. Dzięki temu udało nam się wyprowadzić cztery miary czystości reguł, które służą do sterowania procesem konstrukcji pojedynczej reguły. Przeanalizowaliśmy te miary pod kątem przetargu pomiędzy trafnością predykcji (dyskryminacją) a pokryciem reguły. Wykazaliśmy w eksperymencie obliczeniowym, że zaproponowany algorytm jest konkurencyjny do innych dobrze znanych algorytmów indukcji reguł, takich jak SLIPPER, LRI czy RuleFit. W (Dembczyński i inni, 2010b) omówiliśmy dokładnie różnicę pomiędzy sekwencyjnym pokrywaniem a podejściem zwiększania gradientu. W (Dembczyński i inni, 2010d) rozpatrzyliśmy dwa modele regułowe zastosowane do problemu rangowania. Pierwszy model wykorzystuje funkcję użyteczności, która przypisuje wartość użyteczności do każdego pojedynczego obiektu. Drugi model bazuje na funkcji preferencji zdefiniowanej na parach obiektów. Podczas predykcji otrzymane preferencje na parach są w kolejnym kroku przekształcane do porządku linowego.

5.2 Złożone miary trafności predykcji

W klasyfikacji wieloetykietowej ze względu na wielowymiarowość zmiennej wyjściowej w naturalny sposób pojawiają się złożone funkcje straty. Również w standardowych problemach uczenia maszynowego,

takich jak klasyfikacja binarna, możemy rozważać problem optymalizacji *złożonych miar trafności predykcji*. Standardowa funkcja straty jest zdefiniowana dla pojedynczego przykładu i etykiety. Złożone miary są natomiast obliczane na zbiorze przykładów i/lub etykiet. Nie mogą one zostać w prosty sposób rozłożone na składowe odnoszące się do pojedynczych przypadków i/lub etykiet. Na przykład miara F_β w problemach klasyfikacji binarnej jest obliczana na całym zbiorze testowym. Jej wartość nie może zostać wyrażona jako średnia strata na pojedynczych przykładach testowych. Ponieważ algorytmy nauczone dla jednej funkcji straty mogą być dowolnie słabe ze względu na inną funkcję straty, wydaje się uzasadnione opracowanie algorytmów odpowiednio przystosowanych dla konkretnej (wybranej) funkcji straty. W tym celu warto przeprowadzić analizę teoretyczną miar trafności predykcji z punktu widzenia ich optymalizacji i własności statystycznych.

W pracy (Kotłowski i inni, 2011) dokładnie przeanalizowaliśmy problem dwudzielnego rankingu (ang. *bipartite ranking*), w którym przykłady pozytywne należy uporządkować powyżej przykładów negatywnych. Typową funkcją straty w tym problemie jest błąd rangowy zdefiniowany na parach przykładów. Wykazaliśmy, że ten błąd może być minimalizowany poprzez zastępczą funkcję straty zdefiniowaną na pojedynczych przykładach, np. poprzez stratę logistyczną lub wykładniczą. Główny wynik tej pracy dotyczył ograniczenia żalu błędu rangowego poprzez żal funkcji zastępczej. Ten wynik był dalej przez nas wykorzystany w analizie błędu rangowego w problemie klasyfikacji wieloetykietowej, omówionej powyżej w opisie głównego osiągnięcia naukowego.

Wyniki dla klasyfikacji wieloetykietowej dla różnych funkcji straty zostały przez nas otrzymane w ramach tzw. *podejścia teorio-decyzyjnego*, w którym zakładamy, że zbiór testowy jest stały i skończony (np. zbiór etykiet dla danego przykładu). Natomiast tzw. *podejście populacyjne* definiuje złożone miary trafności predykcji za pomocą wielkości populacyjnych. Dla tego drugiego podejścia przeanalizowaliśmy w pracach (Kotłowski i Dembczyński, 2015, 2017) szeroką rodzinę funkcji liniowo-ułamkowych, zdefiniowaną na macierzy pomyłek (przykładami takich funkcji jest miara F_β, współczynnik Jaccarda oraz wiele innych popularnych miar). Analiza ta dotyczyła dwuetapowej procedury, w której najpierw uczona jest funkcja rzeczywista poprzez minimalizację na zbiorze uczącym błędu zastępczego dla klasyfikacji binarnej (takiego jak błąd kwadratowy, logistyczny lub wykładniczy), a następnie strojony jest próg na osobnym zbiorze walidacyjnym poprzez bezpośrednią optymalizację danej miary trafności predykcji. Udało nam się pokazać, że żal tak otrzymanego klasyfikatora ze względu na daną miarę trafności predykcji jest ograniczony od góry poprzez żal zastępczej funkcji straty. Te wyniki były następnie przez nas rozszerzone do mikro i makro uśrednionych miar trafności predykcji używanych w klasyfikacji wieloetykietowej.

W ostatniej pracy (Dembczyński i inni, 2017) udało nam się dokładnie scharakteryzować dla klasyfikacji binarnej związek pomiędzy podejściem teorio-decyzyjnym a podejściem populacyjnym i wykazać ich asymptotyczną równoważność. Ponadto przedstawiliśmy algorytm przybliżony dla podejścia teoriodecyzyjnego, którego złożoność czasowa jest quasi-liniowa ze względu na liczbę przykładów testowych. Wcześniejsze algorytmy wymagały czasu kwadratowego lub sześciennego.

5.3 Klasyfikacja monotoniczna

W ostatnich latach problem konstrukcji modeli predykcyjnych, które są monotoniczne względem wartości cech, zyskał na znaczeniu zarówno w uczeniu maszynowym, jak i w badaniach operacyjnych. Trudność w zapewnieniu monotoniczności wzrasta wraz ze złożonością trenowanego modelu. Jest to proste zadanie dla modeli liniowych, trochę bardziej złożone w przypadku modeli odcinkami liniowych i regułowych, jednak bardzo trudne dla ogólnej klasy modeli nieliniowych. W pracach (Fallah i inni, 2011, 2012) zaproponowaliśmy wykorzystanie tzw. całki Choqueta do konstrukcji monotonicznych modeli nieliniowych. Choć powszechnie stosowana jako operator agregacji w różnych dziedzinach, np. w wielokryterialnym wspomaganiu decyzji, całka Choqueta jest bardzo mało znana w uczeniu maszynowym. Oprócz łączenia ze sobą takich własności jak monotoniczność oraz elastyczność w elegancki matematyczny sposób, charakteryzuje się ona również dodatkowymi własnościami atrakcyjnymi z punktu widzenia uczenia maszynowego. W szczególności pozwala na określenie w sposób ilościowy ważności pojedynczych cech oraz współzależności pomiędzy grupami cech. Naszym głównym wynikiem było zaproponowanie uogólnienia regresji logistycznej, nazwanego *regresją choqistyczną* (ang. *choquistic regression*), w którym model liniowy został zastąpiony poprzez całkę Choqueta. Wyniki eksperymentalne wykazały, że zaproponowana

metoda jest bardzo konkurencyjna w porównaniu z istniejącymi algorytmami klasyfikacji monotonicznej.

5.4 Predykcja wielu wyjść

Oprócz prac dotyczących klasyfikacji wieloetykietowej stanowiących główne osiągnięcie naukowe, brałem również udział w badaniach nad różnymi jej odmianami. W pracy (Cheng i inni, 2010a) zdefiniowaliśmy problem porządkowej klasyfikacji wieloetykietowej (ang. *graded multi-label classification*), w której zamiast binarnych odpowiedzi możliwe jest udzielenie odpowiedzi na skali porządkowej, określającej stopień przynależności do danej klasy. To rozszerzenie jest motywowane praktycznymi zastosowaniami, w których stopniowana lub częściowa przynależność do klasy jest naturalna. Ten problem może być także spostrzegany jako przykład predykcji z wieloma wyjściami, w którym każde wyjście jest zdefiniowane na skali porządkowej. Zaproponowaliśmy dwa podejścia do rozwiązania tego problemu, które polegają na jego redukcji to standardowego problemu klasyfikacji wieloetykietowej. Ponadto rozważaliśmy różne sposoby mierzenia trafności predykcji w tym problemie.

W pracy (Cheng i inni, 2010b) zaproponowaliśmy dwie nowe metody rozwiązujące problem rangowania etykiet (ang. *label ranking*). W tym problemie predykcja ma postać porządku zupełnego zdefiniowanego na etykietach. Obie zaproponowane metody bazują na probabilistycznym modelu rangowym Placketa-Luce'a. Pierwsza metoda dopasowuje model lokalny oparty na podejściu najbliższych sąsiadów. Druga metoda konstruuje model globalny, którego parametry są reprezentowane jako funkcje obserwacji. Porównując te dwie metody z poprzednimi podejściami, zauważyliśmy że oferują one wiele zalet, na przykład możliwe jest uczenie na podstawie niepełnej informacji o uporządkowaniu etykiet. Ponadto w eksperymencie obliczeniowym pokazaliśmy, że zaproponowane metody uzyskują bardzo konkurencyjne wyniki.

W celu pokazania praktycznego znaczenia naszych badań nad klasyfikacją wieloetykietową ze złożonymi miarami trafności predykcji, wzięliśmy udział w konkursie dotyczącym eksploracji danych zorganizowanym na konferencji *Joint Rough Symposium* w 2012 roku. Konkurs dotyczył klasyfikacji tematycznej artykułów naukowych z zakresu medycyny. Do oceny przesłanych rozwiązań wykorzystana została miara F_1. Nasze rozwiązanie bazujące na algorytmach PCC i GFM zajęło drugie miejsce. Dokładny opis algorytmów i trików zastosowanych przez nas w tym konkursie został opisany w pracy (Cheng i inni, 2012).

5.5 Estymacja czasów przyjazdów w dużej sieci drogowej

We współpracy z firmą NaviExpert, dostarczającej mobilny system nawigacji samochodowej, rozważaliśmy problem estymacji czasów przejazdów w dużej sieci drogowej. W pracy (Dembczyński i inni, 2013b) rozwiązaliśmy ten problem wykorzystując podejście bazujące na faktoryzacji (rozkładzie) macierzy, które jest popularnym narzędziem stosowanych w systemach rekomendacyjnych. Jako macierz wejściową przyjęliśmy macierz czasów przejazdów, w której wiersze odpowiadają krótkim odcinkom drogi, a kolumny 15-minutowym przedziałom czasowym na jakie został podzielony tydzień. Następnie poprzez zastosowanie faktoryzacji macierzy otrzymaliśmy ukryty model cech w postaci dwóch niskowymiarowych macierzy, których iloczyn daje przybliżenie macierzy wejściowej. Otrzymany model charakteryzuje się kilkoma pożądanymi cechami. Zamiast przechowywania pełnej macierzy wejściowej wystarczy zapisać jedynie dwie niskowymiarowe macierze. Estymata czasu przejazdu na danym odcinku drogi i w danym przedziale czasowym może być szybko policzona poprzez pomnożenie odpowiedniego wiersza i kolumny tych macierzy. Ponadto, ukryte cechy otrzymane w wyniku faktoryzacji macierzy dostarczają dodatkowych informacji na temat analizowanego problemu. Eksperyment przeprowadzony na dużym zbiorze danych rzeczywistych pokazał, że zaproponowane podejście jest lepsze od innych modeli szacowania czasu podróży użytych w badaniu.

W pracy (Gawel i inni, 2012b) zajmowaliśmy się problemem zaburzenia estymacji czasu podróży spowodowanego przez nietypową sytuację, taką jak nieoczekiwany korek wynikający z wypadku samochodowego lub awarii drogowej. Zaproponowany przez nas model składa się z dwóch komponentów, statycznego i dynamicznego. Pierwszy z nich odpowiedzialny jest za predykcje długoterminowe, natomiast drugi za predykcje krótkoterminowe. Przetarg pomiędzy obydwoma komponentami jest uzyskiwany

poprzez model regresji liniowej uczony na najnowszych obserwacjach dotyczących ruchu pojazdów. W pracach (Dembczyński i inni, 2012a) oraz (Gawel i inni, 2012a) została przez nas szczegółowo omówiona technologia *community traffic* wykorzystywana przez firmę NaviExpert. Technologia ta pozwala na współpracę społeczności użytkowników systemu w celu poprawy jakości podróżowania. Nasze prace podkreślają, że estymacja czasu podróży, a także optymalizacja proponowanych tras, może być istotnie poprawiona poprzez zbieranie informacji od członków społeczności. Ponadto, bazując na informacji przychodzącej od użytkowników systemu możliwe jest szybkie zlokalizowanie przeszkód drogowych. W tym celu może zostać wykorzystane bezpośrednie wysyłanie informacji o sytuacji drogowej. Do prawidłowego oszacowania prawdopodobieństwa wystąpienia zgłoszonej sytuacji drogowej można wykorzystać dwa podejścia. Pierwsze polega na wykorzystaniu zwykłego głosowania, natomiast drugie wykorzystuje algorytm EM (ang. *expectation-maximization*), który pozwala także na oszacowanie wiarygodności użytkowników zgłaszających sytuacje drogowe.

5.6 Klasyfikacja wiarogodna

Podczas stażu podoktorskiego na Uniwersytecie w Marburgu współpracowałem między innymi z grupą prowadzoną przez prof. Donner-Banzhoffa z Wydziału Medycy Rodzinnej. Głównym celem tej współpracy była analiza bazy danych 1199 pacjentów uskarżających się na bóle w klatce piersiowej. Informacje na temat tych pacjentów zostały zebrane na podstawie badań diagnostycznych z 74 placówek opieki podstawowej. Lekarze ogólni przeprowadzili podstawowy wywiad lekarski oraz standardowe badania medyczne. Zanotowali oni także wstępną diagnozę i zalecone leczenie dotyczące bólów w klatce piersiowej. W analizie opublikowanej w pracy (Hirsch i inni, 2011) wykorzystaliśmy wielowymiarowe metody statystyczne, takie jak analiza korespondencji, wielowymiarowe skalowanie, analiza skupień oraz mapy ciepła, do identyfikacji najważniejszych cech oraz to wyodrębnienia grup pacjentów. Co ciekawe, analiza skupień nie znalazła żadnych wyraźnie wyodrębnionych grup pacjentów. Jako główny wniosek z przeprowadzonej analizy wykazaliśmy, że ból w klatce piersiowej jest heterogeniczną kategorią medyczną bez spójnych powiązań pomiędzy objawami na poziomie pacjenta.

W ramach kontynuacji powyższej współpracy podjęliśmy badania dotyczące właściwej reprezentacji niepewności związanej z predykcją w dziedzinach o szczególnym poziomie bezpieczeństwa, takich jak diagnostyka medyczna. Metody probabilistyczne są często wykorzystywane do modelowania niepewności, jednakże nie rozróżniają one jej dwóch bardzo odmiennych rodzajów: niepewności aleatorystycznej, która wynika ze zmienności statystycznej i ma typową losową naturę, oraz niepewności epistemicznej, która wynika z braku wiedzy. W pracy (Senge i inni, 2014) zaproponowaliśmy metodę *klasyfikacji wiarogodnej* (ang. *reliable classification*), która nie tylko przypisuje etykietę do przykładu, ale także jest w stanie wskazać ilościowo dwa wyżej wymienione źródła niepewności. Metoda ta, pomimo że jest mocno osadzona w teorii prawdopodobieństwa i statystyce, została sformalizowana z wykorzystaniem rozmytych relacji preferencji. Użyteczność zaproponowanego podejścia udało nam się wykazać w eksperymencie obliczeniowym na wyżej wspomnianym zbiorze danych dotyczącym pacjentów z bólem w klatce piersiowej.

Literatura

Agarwal, S. (2014). Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653-1674.

Bartlett, P., Jordan, M., i McAuliffe, J. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138-156.

Beygelzimer, A., Langford, J., i Ravikumar, P. D. (2009). Error-correcting tournaments. W Gavaldà, R., Lugosi, G., Zeugmann, T., i Zilles, S., redaktorzy, *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT 2009)*, wolumen 5809 *Lecture Notes in Computer Science*, strony 247-262. Springer.

Beygelzimer, A., Langford, J., i Zadrozny, B. (2008). Machine learning techniques—reductions between prediction quality metrics. W Liu, Z. i Xia, C. H., redaktorzy, *Performance Modeling and Engineering*, strony 3–28. Springer.

Busa-Fekete, R., Szörényi, B., Dembczyński, K., i Hüllermeier, E. (2015). Online F-measure optimization. W Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., i Garnett, R., redaktorzy, *Advances in Neural Information Processing Systems 28*, strony 595-603. Curran Associates, Inc.

Cheng, W., Dembczyński, K., i Hüllermeier, E. (2010a). Graded multilabel classification: The ordinal case. W Fürnkranz, J. i Joachims, T., redaktorzy, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, strony 223-230. Omnipress.

Cheng, W., Dembczyński, K., i Hüllermeier, E. (2010b). Label ranking methods based on the plackett-luce model. W Fürnkranz, J. i Joachims, T., redaktorzy, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, strony 215-222. Omnipress.

Cheng, W., Dembczyński, K., Hüllermeier, E., Jaroszewicz, A., i Waegeman, W. (2012). F-measure maximization in topical classification. W Yao, J., Yang, Y., Slowinski, R., Greco, S., Li, H., Mitra, S., i Polkowski, L., redaktorzy, *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing, (RSCTC 2012)*, wolumen 7413 *Lecture Notes in Computer Science*, strony 439-446. Springer.

Daumé III, H., Langford, J., i Marcu, D. (2009). Search-based structured prediction. *Machine Learning*, 75:297-325.

Dembczyński, K., Cheng, W., i Hüllermeier, E. (2010a). Bayes optimal multilabel classification via probabilistic classifier chains. W Fürnkranz, J. i Joachims, T., redaktorzy, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, strony 279-286. Omnipress.

Dembczyński, K., Gaweł, P., Jaszkiewicz, A., Kotłowski, W., Kubiak, M., Susmaga, R., Wesołek, P., Wojciechowski, A., i Zielniewicz, P. (2012a). Community traffic: a technology for the next generation car navigation. *Control and Cybernetics*, 41(4):867-883.

Dembczyński, K., Jachnik, A., Kotłowski, W., Waegeman, W., i Hüllermeier, E. (2013a). Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. W Dasgupta, S. i McAllester, D., redaktorzy, *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, wolumen 28 *Proceedings of Machine Learning Research*, strony 1130-1138. PMLR.

Dembczyński, K., Kotłowski, W., Gawel, P., Szarecki, A., i Jaszkiewicz, A. (2013b). Matrix factorization for travel time estimation in large traffic networks. W *Artificial Intelligence and Soft Computing – 12th International Conference (ICAISC 2013)*, wolumen 7895 *Lecture Notes in Computer Science*, strony 500-510. Springer.

Dembczyński, K., Kotłowski, W., i Hüllermeier, E. (2012b). Consistent multilabel ranking through univariate losses. W Langford, J. i Pineau, J., redaktorzy, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, strony 1319-1326. Omnipress.

Dembczyński, K., Kotłowski, W., Koyejo, O., i Natarajan, N. (2017). Consistency analysis for binary classification revisited. W Precup, D. i Teh, Y. W., redaktorzy, *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, wolumen 70 *Proceedings of Machine Learning Research*, strony 961–969. PMLR.

Dembczyński, K., Kotłowski, W., i Słowiński, R. (2010b). Beyond sequential covering – boosted decision rules. W Koronacki, J., Ras, Z. W., Wierzchon, S. T., Kacprzyk, J., i Kacprzyk, J., redaktorzy, *Advances in Machine Learning I*, wolumen 262 *Studies in Computational Intelligence*, strony 209-225. Springer-Verlag.

Dembczyński, K., Kotłowski, W., i Słowiński, R. (2010c). ENDER – a statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery*, 21(1):52-90.

Dembczyński, K., Kotłowski, W., Słowiński, R., i Szeląg, M. (2010d). Learning of rule ensembles for multiple attribute ranking problems. W Fürnkranz, J. i Hüllermeier, E., redaktorzy, *Preference Learning*, strony 217-247. Springer-Verlag.

Dembczyński, K., Kotłowski, W., Waegeman, W., Busa-Fekete, R., i Hüllermeier, E. (2016). Consistency of probabilistic classifier trees. W Frasconi, P., Landwehr, N., Manco, G., i Vreeken, J., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 9852 *Lecture Notes in Computer Science*, strony 511-526. Springer.

Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2010e). Regret analysis for performance metrics in multi-label classification: The case of Hamming and subset zero-one loss. W Balcázar, J. L., Bonchi, F., Gionis, A., i Sebag, M., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 6321 *Lecture Notes in Computer Science*, strony 280-295. Springer-Verlag.

Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2011). An exact algorithm for F-measure maximization. W Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., i Weinberger, K. Q., redaktorzy, *Advances in Neural Information Processing Systems 24*, strony 1404-1412. Curran Associates, Inc.

Dembczyński, K., Waegeman, W., Cheng, W., i Hüllermeier, E. (2012c). On loss minimization and label dependence in multi-label classification. *Machine Learning*, 88:5-45.

Dembczyński, K., Waegeman, W., i Hüllermeier, E. (2012d). An analysis of chaining in multi-label classification. W *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, wolumen 242 *Frontiers in Artificial Intelligence and Applications*, strony 294-299. IOS Press.

Doppa, J., Fern, A., i Tadepalli, P. (2014). Structured prediction via output space search. *Journal of Machine Learning Research*, 15:1317-1350.

Duchi, J. C., Mackey, L. W., i Jordan, M. I. (2010). On the consistency of ranking algorithms. W Fürnkranz, J. i Joachims, T., redaktorzy, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, strony 327-334. Omnipress.

Fagin, R., Lotem, A., i Naor, M. (2003). Optimal aggregation algorithms for middleware. *Journal of Computer and System Sciences*, 66(4):614-656.

Fallah, A., Cheng, W., Dembczyński, K., i Hüullermeier, E. (2011). Learning monotone nonlinear models using the Choquet integral. W Hofmann, T., Malerba, D., Gunopulos, D., i Vazirgiannis, M., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 6913 *Lecture Notes in Computer Science*, strony 414–429. Springer-Verlag.

Fallah, A., Cheng, W., Dembczyński, K., i Hüullermeier, E. (2012). Learning monotone nonlinear models using the Choquet integral. *Machine Learning*, 89(1-2):183-211.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage.

Gao, W. i Zhou, Z.-H. (2013). On the consistency of multi-label learning. *Artificial Intelligence*, 199-200:22-44.

Gawel, P., Dembczyński, K., Kotłowski, W., Kubiak, M., Susmaga, R., Zielniewicz, P., i Jaszkiewicz, A. (2012a). Community traffic: A technology for the next generation car navigation. W Pechenizkiy, M. i Wojciechowski, M., redaktorzy, *New Trends in Databases and Information Systems, Workshop Proceedings of the 16th East European Conference (ADBIS 2012)*, wolumen 185 *Advances in Intelligent Systems and Computing*, strony 339-348. Springer.

Gawel, P., Dembczyński, K., Susmaga, R., Wesolek, P., Zielniewicz, P., i Jaszkiewicz, A. (2012b). Adapting travel time estimates to current traffic conditions. W Pechenizkiy, M. i Wojciechowski, M., redaktorzy, *New Trends in Databases and Information Systems, Workshop Proceedings of the 16th East European Conference (ADBIS 2012)*, wolumen 185 *Advances in Intelligent Systems and Computing*, strony 79-88. Springer.

Hirsch, O., Bösner, S., Hüllermeier, E., Senge, R., Dembczyński, K., i Donner-Banzhoff, N. (2011). Multivariate modeling to identify patterns in clinical data: the example of chest pain. *BMC medical research methodology*, 11(1):155.

Jasinska, K., Dembczyński, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., i Hüllermeier, E. (2016). Extreme F-measure maximization using sparse probability estimates. W Balcan, M. F. i Weinberger, K. Q., redaktorzy, *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, wolumen 48 *Proceedings of Machine Learning Research*, strony 1435-1444. PMLR.

Kotłowski, W. i Dembczyński, K. (2015). Surrogate regret bounds for generalized classification performance metrics. W Holmes, G. i Liu, T.-Y., redaktorzy, *Asian Conference on Machine Learning*, wolumen 45 *Proceedings of Machine Learning Research*, strony 301-316. PMLR.

Kotłowski, W. i Dembczyński, K. (2017). Surrogate regret bounds for generalized classification performance metrics. *Machine Learning Journal*, 106:549-572.

Kotłowski, W., Dembczyński, K., i Hüllermeier, E. (2011). Bipartite ranking through minimization of univariate loss. W Gettor, L. i Scheffer, T., redaktorzy, *Proc. of 28th Annual International Conference on Machine Learning (ICML 2011)*, strony 1113–1120. Omnipress.

Kumar, A., Vembu, S., Menon, A. K., i Elkan, C. (2013). Beam search algorithms for multilabel learning. *Machine Learning*, 92(1):65-89.

Kurzynski, M. (1988). On the multistage bayes classifier. *Pattern Recognition*, 21(4):355-365.

Lafferty, J. D., McCallum, A., i Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. W Brodley, C. E. i Danyluk, A. P., redaktorzy, *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, strony 282–289. Morgan Kaufmann.

McAllester, D. (2009). Generalization bounds and consistency for structured labeling. W *Predicting Structured Data*, strony 247–261. MIT Press.

McCallum, A., Freitag, D., i Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. W Langley, P., redaktor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, strony 591-598. Morgan Kaufmann.

Morin, F. i Bengio, Y. (2005). Hierarchical probabilistic neural network language model. W Cowell, R. G. i Ghahramani, Z., redaktorzy, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, strony 246-252. Society for Artificial Intelligence and Statistics.

Petterson, J. i Caetano, T. S. (2010). Reverse multi-label learning. W Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., i Culotta, A., redaktorzy, *Advances in Neural Information Processing Systems 23*, strony 1912-1920. Curran Associates, Inc.

Petterson, J. i Caetano, T. S. (2011). Submodular multi-label learning. W Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., i Weinberger, K. Q., redaktorzy, *Advances in Neural Information Processing Systems 24*, strony 1512-1520. Curran Associates, Inc.

Platt, J. C., Cristianini, N., i Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. W Solla, S. A., Leen, T. K., i Müller, K., redaktorzy, *Advances in Neural Information Processing Systems 12*, strony 547–553. MIT Press.

Prabhu, Y. i Varma, M. (2014). FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. W Macskassy, S. A., Perlich, C., Leskovec, J., Wang, W., i Ghani, R., redaktorzy, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014)*, strony 263-272, New York, NY, USA. ACM.

Read, J., Pfahringer, B., Holmes, G., i Frank, E. (2009). Classifier chains for multi-label classification. W Buntine, W. L., Grobelnik, M., Mladenic, D., i Shawe-Taylor, J., redaktorzy, *Machine Learning and Knowledge Discovery in Databases*, wolumen 5782 *Lecture Notes in Computer Science*, strony 254-269. Springer.

Read, J., Pfahringer, B., Holmes, G., i Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333-359.

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., i Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16-29.

Stock, M., Dembczyński, K., Baets, B. D., i Waegeman, W. (2016). Exact and efficient top-k inference for multi-target prediction by querying separable linear relational models. *Data Mining and Knowledge Discovery*, 30(5):1370-1394.

Tewari, A. i Bartlett, P. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007-1025.

Tsochantaridis, Y., Joachims, T., Hofmann, T., , i Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453-1484.

Waegeman, W., Dembczyński, K., Jachnik, A., Cheng, W., i Hüllermeier, E. (2014). On the Bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research*, 15(1):3333-3388.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., i Attenberg, J. (2009). Feature hashing for large scale multitask learning. W Danyluk, A. P., Bottou, L., i Littman, M. L., redaktorzy, *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, strony 1113-1120. ACM.

Ye, N., Chai, K. M. A., Lee, W. S., i Chieu, H. L. (2012). Optimizing F-measure: A tale of two approaches. W Langford, J. i Pineau, J., redaktorzy, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, strony 289-296. Omnipress.

Krzysztof Dembczyński