4D1-4in2

重症度を考慮した医学単語重み付け手法による死亡予測

Mortality prediction using a semantic term weighting method based on severity of patients

松尾亮輔　　Ho Tu Bao
Ryosuke Matsuo　　Tu Bao Ho

北陸先端科学技術大学院大学知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Science and Technology

Term weighting is essential in representing documents by term vectors for text classification. Whereas the term frequencies such as TFIDF and its variants have often been employed in term weighting, semantic term weighting is necessary for medical domain where ontologies are commonly employed to capture the semantics of medical terms. However, medical importance of terms from the viewpoint of severity of patients is still not identified by ontologies. To capture the medical importance such as the severity of patients by utilizing ontologies, we propose a semantic term weighting method that combines mapping information from the medical ontology UMLS, the hierarchical structure of ICD-10, and the linear relationship of medical terms from a ranking of causes of death. The results of the proposed method showed high performance in mortality prediction. The proposed method captures the severity of patients from various medical terms and can be applied to the patient risk prediction.

1. はじめに

単語の重み付けは文書内で単語の重要度を区別する手法であり，単語ベクトルでの文書表現が可能となるため，文書分類やクラスタリング，センチメント分析などに適用されている．単語に重み付けをする際によく用いられる手法は TFIDF [Salton 88] で，この手法は TF (term frequency) という単語の出現頻度と，IDF (Inverse document frequency) という逆文書頻度の 2 つの統計的指標を用いて単語の重要度を決定して重み付けをする．TFIDF は最新の手法ではないがシンプルかつ効果的であることから，新たなアルゴリズムを形成するためのベースとして良く用いられている [Ramos 03].

この TFIDF による単語の重み付けは単語の頻度情報から単語の重要度を捉えることが出来るが，単語の意味的側面からはその重要度を捉えていない．この制限を乗り越えるために，オントロジーを使った単語の重み付け手法が開発されている．医学文書に適用した手法として，1 つはカテゴリーとの意味的な類似度をもとに単語の重みを与えている [Luo 11]．この手法は WordNet というオントロジーを用いている．その WordNet 内の数あるカテゴリーの 1 つに医学ドメインがあり，ある単語が WordNet に存在する場合にそのカテゴリーと単語の類似度を計算する．そのほか，医学オントロジーである MeSH オントロジーを使って単語間の意味的な関係性を考慮した重み付け手法がある [Zhang 07, Zhang 08]．この手法は同じベクトル内に意味的に似た医学単語が発生した場合にそれらの単語の重みを上昇させる．また UMLS (Unified Medical Language System) という医学オントロジーを用いることで単語の意味を考慮した重み付け手法がある [Yu 09]．この手法は UMLS の概念情報を活用してクエリー内の単語を UMLS concept や UMLS synonym というようにカテゴリー分けをし，そのカテゴリーをもとに重みを上昇させている．

このようにオントロジーを活用して医学単語に対してその意味を考慮した重み付け手法は多くあるが，患者の重症度という視点から医学単語の意味的な重要度は捉えられていない．

連絡先: 松尾亮輔，北陸先端科学技術大学院大学知識科学研究科，石川県能美市旭台 1 − 1，matsuor@jaist.ac.jp

そこで本研究は医学オントロジーである UMLS を活用しながら，患者の重症度という観点から医学単語の重要度を捉える．そのため，本研究では UMLS から得られるマッピング情報と ICD-10 の階層構造，死因ランキングによる医学単語間の線形関係を組み合わせることで，患者の重症度を考慮した医学単語の重み付け手法を提案する．

2. 提案手法

本研究では，ICD-10 に基づいて患者の重症度を考慮した単語の医学重要度を捉える．ICD (International Statistical Classification of Diseases and Related Health Problems) は診断用の医学単語の英数字コードを提供しており，元は死亡原因を分類するためのもので死亡率のデータを処理するために適用されてきている [Organization 04]．したがって ICD-10 は患者の重症度を捉えるのに適していると考えられる．

本提案手法はその ICD-10 を軸に主に 3 つのステップによって構成される．はじめに医学文書内の単語と ICD-10 単語との関連づけをする．その後，ICD-10 コードを活用して重症度を考慮した医学重要度の程度を識別し，最後に医学重要度の重みと TFIDF の重みを組み合わせることで提案手法の重みを求める．以下にこれら 3 つのステップの詳細を順に述べる．

2.1 医学文書内の単語と ICD-10 単語の関連づけ

このステップでは，UMLS (Unified Medical Language System) [Bodenreider 04] の概念情報を活用して，医学文書内の単語と ICD-10 単語の関連づけを行う．まず MetaMap [Aronson 01] を用いて UMLS concept を医学文書内の単語から識別する．そしてある単語が UMLS concept を持っている場合，その単語を医学単語として扱う．その後，UMLS concept を持つ医学単語に対して，BioPortal [Noy 09] を活用することで，UMLS から得られる医学単語の概念情報をもとに，ICD-10 単語との関連づけを試みる．BioPortal 上で関連づけが出来た場合，その医学単語を ICD-10 単語として扱う．このような手順を経ることで，医学文書内の単語が ICD-10 単語へと関連づけられる．ICD-10 単語の ICD-10 コードは次のステップで活用する．

2.2 重症度を考慮した医学重要度合いの識別

前のステップで文書内の ICD-10 単語を捉えることが出来た．本ステップでは，ICD-10 単語の重症度を考慮した医学重要度の程度を識別するために，トップ 15 の死因ランキング [Murphy 13] を活用する．このランキングを用いることで，線形関係に沿って実数で ICD-10 単語に対して患者の重症度の程度を付与することが出来るようになる．このランキングはそれぞれのランクを死因名とそれに対応する ICD-10 コードで表現している．そのため，ICD-10 単語の ICD-10 コードを利用することで，それぞれのランクへのひも付けが可能となり，そのランク情報により ICD-10 単語の重症度の程度を識別出来る．本研究では，以下の式 (1) を用いてランク情報から重症度を考慮した医学重要度の程度を計算する．

$$v_i = \frac{(v_{max} - v_{min}) \times (\xi - i + 1)}{\xi} \quad (1)$$

v_i は順位 i 番目のランクに対応する医学重要度である．v_{max} と v_{min} は医学重要度の最大値と最小値にあたり，本研究ではそれぞれ 0.9，0.2 と設定する．ξ はランクにより区分けされるグループの総数を意味する．まずトップ 15 の死因ランキングから 15 のグループをランク順位に沿って設定出来る．しかしそのランキングでは対応出来ない医学単語があるため，本研究ではそのランキングに該当しないが ICD-10 単語である医学単語を 16 番目のグループ，ICD-10 単語ではない医学単語を 17 番目のグループとする．よって，ここで用いる ξ の値は 17 となる．以下の表 1 は死因ランキングの死因名とそれに対応する ICD-10 コード，医学重要度の値を表している．表 1 には記載されていないが，死因ランキング外の 16 番目のランクグループと，17 番目のランクグループも同じように式 (1) を用いて計算し，医学重要度はそれぞれ 0.08，0.04 となる．

表 1: トップ 15 の死因ランキング [Murphy 13] の死因名とそれに対応する ICD-10 コード及び医学重要度の値

The rank	The name of cause of death	The ICD-10 code(s)	The value
1	Disease of heart	I00-I09, I11, I13, I20-I51	0.7
2	Malignant neoplasms	C00-C97	0.66
3	Chronic lower respiratory diseases	J40-J47	0.62
4	Cerebrovascular diseases	I60-I69	0.58
5	Accidents (unintentional injuries)	V01-X59, Y85-Y86	0.54
6	Alzheimer's disease	G30	0.49
7	Diabetes mellitus	E10-E14	0.45
8	Nephritis, nephritic syndrome and nephrosis	N00-N07, N17-N19, N25-N27	0.41
9	Influenza and pneumonia	J09-J18	0.37
10	Intentional self-harm (suicide)	U03, X60-X84, Y87.0	0.33
11	Septicemia	A40-A41	0.29
12	Chronic liver disease and cirrhosis	K70, K73-K74	0.25
13	Essential hypertension and hypertensive renal disease	I10, I12, I15	0.21
14	Parkinson's disease	G20-G21	0.16
15	Pneumonitis due to solids and liquids	J69	0.12

また，ICD-10 は階層構造であるため，その特徴を活用して死因ランキングのランクに対応する ICD-10 単語の医学重要度の重みをその下位に位置する ICD-10 単語へ伝搬させる．たとえば，malignant neoplasm は死因ランキングでランク 2 の医学単語であり，ICD-10 コードでは C00-C97 の範囲の ICD-10 単語に該当する．その際，ICD-10 コード C00-C97 の ICD-10 単語だけでなく，それらの ICD-10 コードの下位に属する ICD-10 単語に対しても医学重要度として 0.66 の重みを同じように付与する．このように ICD-10 の階層構造から階層内の単語の意味的な類似性を考慮することにより，より多くの医学単語を重み付けの対象とすることを可能にする．

2.3 重みの組み合わせ

最終的に前のステップで求めた重症度を考慮した医学重要度の重み（MED）は，Min-Max 正規化後の TFIDF の重み（TFIDF）と組み合わせることで提案手法の重みとなる．組み合わせ方は以下の式 (2) となる．

$$Proposed\ weight = \alpha^1 * TFIDF * \{1 + (\alpha^2 * MED)\} \quad (2)$$

ここで，α^1 と α^2 は TFIDF と MED の重みの係数で，それぞれ 0.5，1.5 とする．TFIDF の影響を小さくし，医学重要度の影響を大きくするために上記の係数を用いる．

3. 実験結果

本研究の提案手法の有効性を見るために死因予測の実験を行う．データセットは MIMIC II データベースから 60 歳以上の患者の電子カルテで合計 13,026 文書を用いる．ストップワードの除去とチャンキングを前処理として行い，13,026 文書からそれぞれの文書内の単語の TFIDF の重みを計算する．予測のための入力データでは 20 以上の文書で出現した単語のみを扱い，その単語数は 11,858 である．

死因予測の際には 2 つのラベルを用いる．ラベル 1 は患者が病院で亡くなった場合でラベル 0 は亡くならなかった場合である．この 2 つのラベルは MIMIC II データベースの expire_flg のカラムから取得する．ラベル 1 とラベル 0 の文書数はそれぞれ 2,158，10,868 である．この実験では，2 つのラベルに該当する文書数の割合を一定にし，それぞれのラベルごとに 500 文書で合計 1,000 文書から死因予測を実行する．

提案手法は以下の 3 つの手法と比較される．1 つ目はベースラインとして用いる TFIDF（TFIDF）である．2 つ目は UMLS により識別された医学単語をすべて同じ重みで上昇させた手法（TFIDF-MED-1）である．この場合，医学重要度は一定である．3 つ目は死因分類として使われている ICD-10 コードを持つ医学単語に ICD-10 コードを持たない医学単語よりも高い重みを与えた手法（TFIDF-MED-2）である．この場合は 2 つのパターンで医学重要度を捉える．提案手法はさらに ICD-10 コードを持つ医学単語に対して死因ランキングのランク順位に基づいて重みを上昇させた手法（TFIDF-MED-Ranking）である．TFIDF-MED-1 と TFIDF-MED-2 は提案手法である TFIDF-MED-Ranking と同じように係数を用いて TFIDF と医学重要度の重みを組み合わせる．

表 2: 累積平均による精度の比較

The range	TFIDF	TFIDF-MED-1	TFIDF-MED-2	TFIDF-MED-Ranking
1-50	0.8255	0.8257	0.826	**0.8284**
1-100	0.8226	0.8232	0.8233	**0.8252**
1-150	0.8171	0.818	0.818	**0.8203**
1-200	0.8186	0.8191	0.8191	**0.8209**
1-250	0.8203	0.8207	0.8207	**0.8225**
1-300	0.8191	0.8196	0.8197	**0.8218**
1-350	0.8202	0.8208	0.821	**0.8231**
1-400	0.8193	0.8198	0.82	**0.8223**
1-450	0.8199	0.8205	0.8207	**0.8229**
1-500	0.8191	0.8196	0.8198	**0.8223**

表 3: 平均精度の比較

The range	TFIDF	TFIDF-MED-1	TFIDF-MED-2	TFIDF-MED-Ranking
1-50	0.8255	0.8257	0.826	**0.8284**
50-100	0.8213	0.8222	0.8221	**0.8234**
100-150	0.8073	0.8084	0.8086	**0.8113**
150-200	**0.8215**	0.8213	0.8209	**0.8215**
200-250	0.8281	0.8279	0.828	**0.8295**
250-300	0.8121	0.8129	0.814	**0.8176**
300-350	0.8278	0.8287	0.8291	**0.8317**
350-400	0.8121	0.8132	0.8131	**0.8163**
400-450	0.8237	0.8249	0.8251	**0.8265**
450-500	0.813	0.8123	0.8135	**0.818**

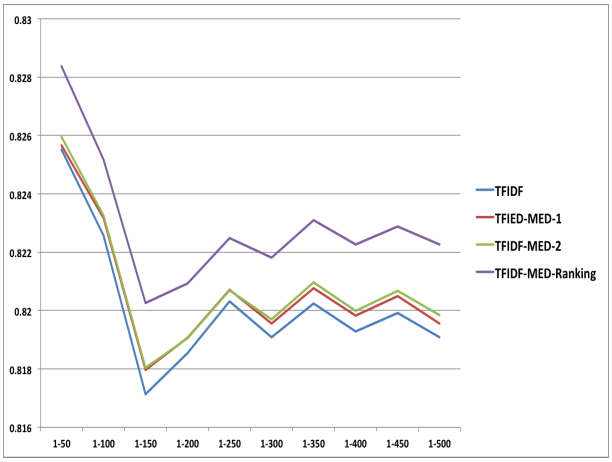

図 1: 累積平均による精度の推移グラフ

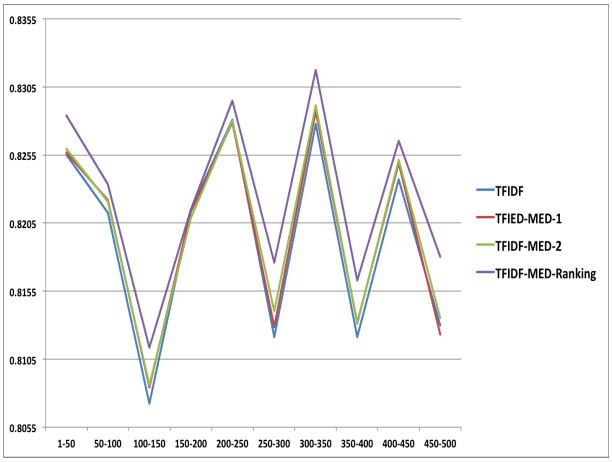

図 2: 平均精度の推移グラフ

本研究では，Support vector machine（RBF カーネル）を用いて死因予測を行う．特徴選択は L2 正則化を用いる．これらは Scikit-learn [Pedregosa 11] のライブラリを使って実行する．また，本研究の実験では RBF カーネルと L2 正則化のパラメータはデフォルトとする．そして 1,000 の文書をラベルの割合を一定にしながらランダムに変更して，5-fold cross-validation を 500 回実施することで提案手法の有効性を調べる．

表 2 と図 1 は 5-fold cross-validation により得られた精度の累積平均の結果である．この結果から提案手法である TFIDF-MED-Ranking が他の 3 つの手法と比べて全ての条件で高い精度を示していることがわかる．表 3 と図 2 は 50 回ごとの 5-fold cross-validation の精度の平均値である．提案手法（TFIDF-MED-Ranking）は 150 回目から 200 回目までの精度の平均値以外の全ての場合で他の手法よりも精度が良いことがわかる．

4. 考察と結論

提案手法は患者の重症度を考慮した医学重要度の重みを捉えるために，UMLS， BioPortal，ICD-10，死因ランキングといった医学知識を用いているため，提案手法による結果は医学分野の知識に従って導かれている．

その提案手法（TFIDF-MED-Ranking）は死因予測の実験結果から比較した他の手法と比べて高い精度を示した．特徴選択として用いた L2 正則化と Support vector machine（RBF カーネル）への入力ベクトルを ICD-10 を軸に患者の重症度を考慮しながら変更したことが精度の向上に貢献したと考えられる．特に医学重要度を一定にした場合の手法である TFIDF-MED-1 と 2 つのパターンで医学重要度を捉えた場合の手法である TFIDF-MED-2 よりも，死因ランキングのランク情報に基づいて患者の重症度の程度を捉えた提案手法が良い精度をもたらしたことから，死因ランキングによる医学重要度の可変の有効性が示されたと考えられる．

提案手法の特徴は UMLS から得られるマッピング情報や，ICD-10 の階層構造，死因ランキングの線形関係といったように，医学単語間の様々な種類の関係性を組み合わせて重み付けをしている点である．また，ある特定の病気だけではなく，様々な病気に対する患者の重症度を捉えている点も提案手法の特徴である．提案手法による患者の重症度を考慮した重み付けは実験結果から死因予測のタスクで有効なことが分かったため，患者のリスク予測問題についても本提案手法が適用出来ると考えられる．今後は，ある ICD-10 単語の下位に位置する複数の ICD-10 単語間の医学重要度の差異を考慮することと，死因ランキング外の ICD-10 単語や医学単語に対する医学重要度の重みの調整を行う．

5. 謝辞

本研究は，JAIST Data Science Project と Vietnam National University at Ho Chi Minh City under the grant number B2015-42-02 から部分的な助成を受けて行われた．

参考文献

[Aronson 01] Aronson, A. R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., in *Proceedings of the AMIA Symposium*, p. 17American Medical Informatics Association (2001)

[Bodenreider 04] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research*, Vol. 32, No. suppl 1, pp. D267–D270 (2004)

[Luo 11] Luo, Q., Chen, E., and Xiong, H.: A semantic term weighting scheme for text categorization, *Expert Systems with Applications*, Vol. 38, No. 10, pp. 12708–12716 (2011)

[Murphy 13] Murphy, S. L., Xu, J., and Kochanek, K. D.: Deaths: final data for 2010., *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, Vol. 61, No. 4, pp. 1–117 (2013)

[Noy 09] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research*, p. gkp440 (2009)

[Organization 04] Organization, W. H.: *International statistical classification of diseases and related health problems*, Vol. 1, World Health Organization (2004)

[Pedregosa 11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011)

[Ramos 03] Ramos, J.: Using tf-idf to determine word relevance in document queries, in *Proceedings of the first instructional conference on machine learning* (2003)

[Salton 88] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information processing & management*, Vol. 24, No. 5, pp. 513–523 (1988)

[Yu 09] Yu, H. and Cao, Y.-G.: Using the weighted keyword models to improve information retrieval for answering biomedical questions, *AMIA summit on translational bioinformatics* (2009)

[Zhang 07] Zhang, X., Jing, L., Hu, X., Ng, M., and Zhou, X.: A comparative study of ontology based term similarity measures on PubMed document clustering, in *Advances in Databases: Concepts, Systems and Applications*, pp. 115–126, Springer (2007)

[Zhang 08] Zhang, X., Jing, L., Hu, X., Ng, M., Jiangxi, J. X., and Zhou, X.: Medical document clustering using ontology-based term similarity measures, *International Journal of Data Warehousing and Mining (IJDWM)*, Vol. 4, No. 1, pp. 62–73 (2008)