Relevance Assessment of Clinical Statements

Michel Oleynik

Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz, Austria

2016

Motivation

- Huge amount of clinical data produced over the years
- Different content types
  - Unstructured and semi-structured texts
  - Numeric and coded data
  - Biosignals and images
- Clinical text
  - Short forms: abbreviations, acronyms
  - Spelling and typing errors
  - Short, incomplete sentences
- Two problems addressed
  - Unsupervised expansion of ad-hoc abbreviations in EHR narratives
  - Automated classification of pathology reports into ICD-O codes

Outline

Goal

Original

3. St.p. TE eines exulz. sek.knot.SSM (C43.5) li Lab. majus.
Level IV, 2,42 mm Tumordurchm.

Goal

Original

3. St.p. TE eines exulz. sek.knot.SSM (C43.5) li Lab. majus.
Level IV, 2,42 mm Tumordurchm.

Expanded

Status post Totalexzision eines exulzerierenden sekundär knotigen superfiziell spreitenden Melanoms (C43.5) linkes Labium majus.
Level Vier, 2,42 Millimeter Tumordurchmesser.

[English translation]

3. History of total excision of an exulcerated secondarily nodular superficially spreading melanoma (C43.5) of the outer left labia.
Level 4, tumor diameter 2.42mm.

Materials and Methods

- 30,000 clinical documents from the cardiology domain
  - Written in German by Austrian physicians
  - Discharge summaries, finding reports
  - Routine documentation in LKH, Graz
- Evaluation data
  - Random subsequences of 100 characters
  - 147 abbreviations manually expanded by a human annotator
- Unsupervised machine learning
  - Bigram and unigram lookup
  - Training-test split (90% - 10%)
  - Evaluated by accuracy

Materials and Methods

Figure: Bigram lookup.

Partial Results[1]

Table: Precision, recall and F-score in different matching strategies.

Matching	**N-Gram**	**P**	**R**	F_1
Relaxed	1	0.62	0.62	0.62
Strict	1	0.76	0.71	0.74
Relaxed	2	0.91	0.81	0.86
Strict	2	0.94	0.73	0.82
Relaxed	3	0.63	0.16	0.26
Strict	3	0.91	0.20	0.32
Combined	-	**0.93**	**0.93**	**0.93**

[1] Submitted to the 16th World Congress on Medical and Health Informatics [Medinfo 2017]

Outline

Goal

Hospital A.C.Camargo
Centro de Tratamento, Ensino e Pesquisa em Câncer

Prontuário: 10304520
Paciente: [redacted]
Convênio: [redacted]
Médico Solicitante: MAURICIO DOI
Setor: MASTOLOGIA

Atendimento: 1747072 E
Pedido de Exame: 579591
Laudo Interno: 1032588
Data: 29/10/2010
Idade: [redacted]

EXAME ANATOMOPATOLÓGICO

REVISÃO DE LÂMINAS

Recebemos para revisão 01 bloco e 01 lâmina identificados como M10-20990, provenientes do Serviço de Anatomia Patológica de Santos, acompanhados do respectivo laudo anátomo-patológico. Aqui, foram identificados como A-1032588.

. Biópsia de mama direita:

* Carcinoma ductal invasivo
* Grau nuclear: 2
* Grau I de SBR
* Mitoses: 0-1/10 CGA
* Desmoplasia peritumoral: Intensa
* Infiltrado linfocitário peritumoral: Discreto
* Invasão vascular linfática: Não detectada
* Invasão vascular sanguínea: Não detectada
* Invasão perineural: Não detectada
* Necrose: Ausente

Nota: Foi solicitada a realização de estudo imunoistoquímico para complementação diagnóstica.

Figure: Example of a pathology report mapped into ICD-O codes.

Materials and Methods

- 70,000+ documents from 20,000+ patients over 14 years
  - Written in Portuguese by Brazilian pathologists
  - Pathology reports
  - Routine procedure in A.C. Camargo Cancer Center, Brazil
- Gold standard creation
  - Free-text content associated to structured data in cancer registries
  - Discarded patients with confirmed metastasis or multiple classifications
- Supervised machine learning
  - Micro-averaged with 10-fold cross-validation
  - Evaluated by precision, recall and F-score
  - Support vector machines (SVMs) with *tf-idf* weighting scheme

$$f(\vec{x}) = sign(\vec{w}^T \vec{x} + b) \quad (1)$$

Materials and Methods

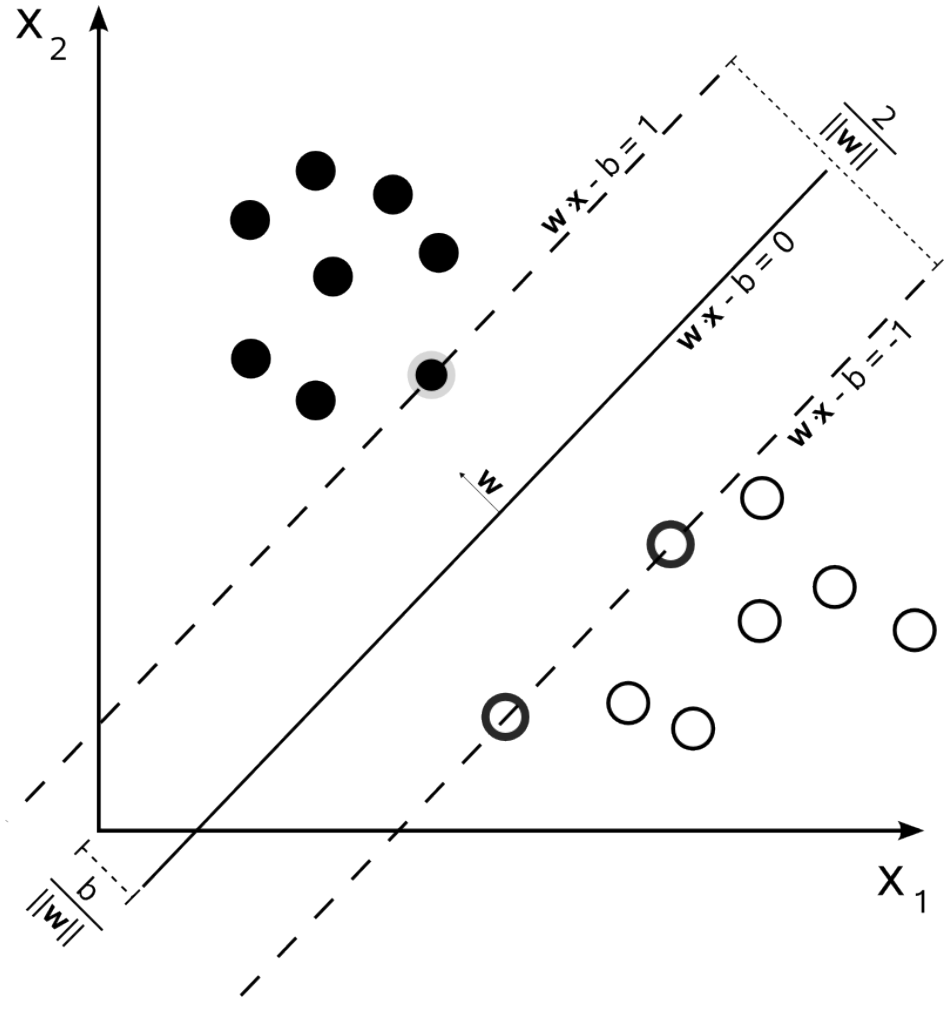

Figure: A support vector machine. [Public domain]

Some Results[2]

Table: Top ten F_1-scores in the ICD-O topography attribution task.

Code Group	Description	n	P	R	F_1
C44	Skin	3,858	0.88	0.94	0.91
C50	Breast	3,668	0.89	0.91	0.90
C73-C75	Thyroid and other endocrine glands	1,329	0.92	0.87	0.90
C60-C63	Male genital organs	1,536	0.93	0.81	0.87
C64-C68	Lymph nodes	660	0.86	0.78	0.82
C51-C58	Female genital organs	1,574	0.85	0.77	0.81
C69-C72	Eye, brain and other parts of central [...]	536	0.83	0.70	0.76
C00-C14	Lip, oral cavity and pharynx	903	0.80	0.71	0.75
C15-C26	Digestive organs	2,159	0.67	0.84	0.75
C77	Lymph nodes	590	0.68	0.80	0.74
	Overall	**18,905**	**0.82**	**0.82**	**0.82**

[2]Submitted to the Medical Informatics Europe Conference [MIE 2017]

Final Remarks

- Contribution
  - Easy process for obtaining structured information over textual data
  - Accelerates the work of physicians when classifying patient data
  - Allows cross-patient search: cohort building for clinical trials
- Future work
  - Assess the *kappa* factor among specialists
  - Evaluate in other domains and languages
  - Relevance assessment of clinical statements

Thank you!

- Contact
  - Michel Oleynik
  - michel.oleynik@stud.medunigraz.at
- Acknowledgments
  - Brazilian National Research Council (CNPq)
  - Medical University of Graz (Med Uni Graz)
  - JULIE Lab, University of Jena, Germany
  - Steiermärkische Krankenanstaltengesellschaft (KAGes)
  - Center for Biomarker Research in Medicine (CBmed)

Additional Results[3]

Table: Top ten F_1-scores in the ICD-O morphology attribution task.

Code Group	Description	n	P	R	F_1
959-972	Hodgkin and non-Hodgkin lymphomas	859	0.85	0.87	0.86
850-854	Ductal and lobular neoplasms	3,410	0.85	0.87	0.86
855	Acinar cell neoplasms	1,059	0.87	0.85	0.86
809-811	Basal cell neoplasms	1,704	0.80	0.89	0.84
872-879	Nevi and melanomas	1,473	0.87	0.81	0.84
906-909	Germ cell neoplasms	208	0.89	0.71	0.79
812-813	Transitional cell papillomas and carcinomas	384	0.81	0.74	0.78
938-948	Gliomas	237	0.82	0.71	0.76
858	Thymic epithelial neoplasms	17	1.00	0.59	0.74
868-871	Paragangliomas and glomus tumors	26	1.00	0.58	0.73
	Overall	**18,599**	**0.74**	**0.74**	**0.73**

[3]Submitted to the Medical Informatics Europe Conference [MIE 2017]