Элементы теории свидетельств

Примеры задач анализа экспертной информации

Задача 1. Предположим, что 10 экспертов дают прогноз о перспективности развития трех технологий $\{a,b,c\}$:

$$3 - \{a, b\}, \qquad 2 - \{b\},$$
$$4 - \{b, c\}, \qquad 1 - \{c\}.$$

Необходимо дать оценки вероятностям перспективности развития каждой технологии.

Задача 2. Предположим, что другая группа из 5-ти экспертов независимо от первой группы (задача 1) дала следующие оценки:

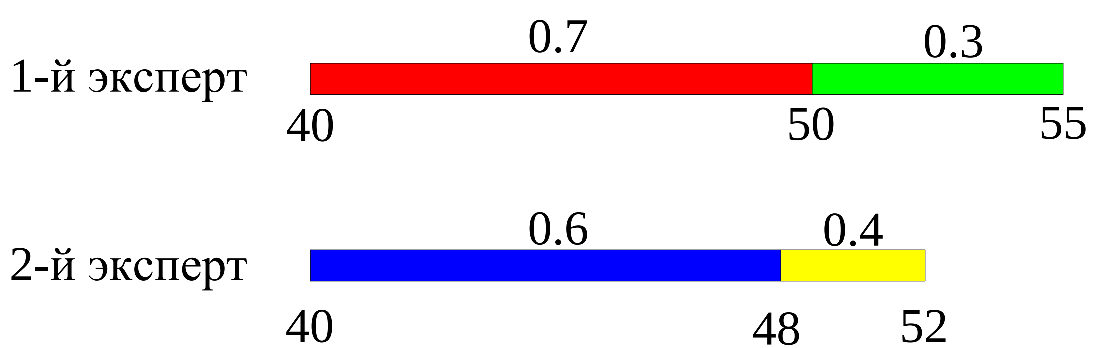
$$2 - \{a, c\}, \qquad 3 - \{b\}.$$

Необходимо: 1) оценить конфликтность информации от двух групп экспертов; 2) оценить и сравнить степени информационной неопределенности двух групп экспертных оценок; 3) построить агрегированные оценки.

Примеры задач

Задача 3. Предположим известным, что в задаче 2 надежность информации от первой группы экспертов, выше надежности информации от второй группы экспертов на p=20%. Как изменятся в этом случае решения задач 2?

Задача 4. Пусть два эксперта независимо дают следующую информацию о прогностической стоимости акций некоторой компании через месяц



Вопросы: 1) насколько конфликтны эти два прогноза? 2) как можно корректно агрегировать эти прогнозы в один прогноз?; 3) тот же вопрос, если эксперты имеют разные «надежности».

Теория свидетельств

Особенности этих задач: малые выборки, нет информации о зависимости экспертных оценок, нет информации о распределениях «внутри» множеств-оценок.

= теория свидетельств (theory of evidence),
= теория функций доверия (theory of belief function),
= теория случайных множеств (theory of random set),
= теория Демпстера – Шейфера (Dempster – Shafer theory)

Dempster A.P. Upper and Lower Probabilities Induced by a Multivalued Mapping// The Ann. of Math. Stat. 1967, 38(2), 325–339.

Shafer G. A Mathematical Theory of Evidence. – Princeton, N.J.: Princeton University Press, 1976.

Основные понятия теории свидетельств

- X – универсальное множество всех возможных значений результатов эксперимента (наблюдений, альтернатив и т.д.);
- 2^X – множество всех подмножеств из X;
- **функция масс (базовая вероятность)** $m\colon 2^X \to [0,1]$, $m(\varnothing)=0$ и $\sum_{A\in 2^X} m(A)=1$ (конечное число ненулевых слагаемых).

Н/р, $m(A)=c/N$, где c – количество наблюдений $x\in A$, N – общее количество наблюдений.

- **фокальный элемент** $A\subseteq X$, если $m(A)>0$;
  $\mathcal{A}=\{A\}$ – множества всех фокальных элементов;
- **тело свидетельства** $F=(\mathcal{A},m)$, где $m(A)$, $A\in\mathcal{A}$;
  $\mathcal{F}(X)$ – множество всех тел свидетельств на X.
- **категоричное** (categorical) тело свидетельств $F_A=(A,1)$;
  н/р, стоимость акций будет в интервале $A=[40,50]$;
- **бессодержательное** (vacuous) тело свидетельств $F_X=(X,1)$;
  н/р, стоимость акций будет в интервале $X=[0,+\infty)$;

Функции доверия и правдоподобия

Функция доверия (belief function): $Bel(A) = \sum_{B \subseteq A} m(B)$,

$$m(A) = \sum\nolimits_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B)$$

(преобразование Мёбиуса для конечного множества X)

Функция правдоподобия (plausibility function): $Pl(A) = \sum_{B \cap A \neq \varnothing} m(B)$.

Отношение двойственности: $Pl(A) + Bel(\neg A) = 1$.

J. Halpern & R. Fagin (1992): функцию доверия можно определить, как функцию, удовлетворяющую ослабленным вариантам аксиом Колмогорова, характеризующих вероятность.

$\Rightarrow$ *Bel(A)*, *Pl(A)* – нижняя и верхняя оценки вероятности события *A*:

$$Bel(A) \leq Pr(A) \leq Pl(A).$$

Задача 1. Точечное решение

Предположим, что 10 экспертов дают прогноз о перспективности развития трех технологий $\{a,b,c\}$:

$$3 - \{a, b\}, \qquad 2 - \{b\},$$

$$4 - \{b, c\}, \qquad 1 - \{c\}.$$

Необходимо дать оценки вероятностям перспективности развития каждой технологии.

Решение 1. Считая, что в случае, когда эксперты высказываются в пользу нескольких альтернатив (технологий), то их оценки равномерно распределены между отдельными технологиями (**сильное предположение!**), получим точечные оценки указанных вероятностей:

$$\tilde{P}(\{a\}) = \frac{1.5}{10} = \frac{3}{20}, \quad \tilde{P}(\{b\}) = \frac{1.5+2+2}{10} = \frac{11}{20}, \quad \tilde{P}(\{c\}) = \frac{2+1}{10} = \frac{3}{10}.$$

Задача 1. Интервальное решение

Если не предполагать известным распределение между альтернативами, то функции масс равны

$$m(\{a,b\})=\frac{3}{10},\ m(\{b,c\})=\frac{4}{10},\ m(\{b\})=\frac{2}{10},\ m(\{c\})=\frac{1}{10}.$$

Тогда

$$Bel(\{a\})=0, \qquad Pl(\{a\})=1-Bel(\{b,c\})=1-0.7=0.3,$$
$$Bel(\{b\})=0.2, \qquad Pl(\{b\})=1-Bel(\{a,c\})=1-0.1=0.9,$$
$$Bel(\{c\})=0.1, \qquad Pl(\{c\})=1-Bel(\{a,b\})=1-0.5=0.5,$$
$$Bel(\{a,b\})=0.3+0.2=0.5, \qquad Pl(\{a,b\})=1-Bel(\{c\})=1-0.1=0.9,$$
$$Bel(\{a,c\})=0.1, \qquad Pl(\{a,c\})=1-Bel(\{b\})=1-0.2=0.8,$$
$$Bel(\{b,c\})=0.4+0.2+0.1=0.7, \qquad Pl(\{b,c\})=1-Bel(\{a\})=1-0=1.$$

Следовательно,

$$0\le P(\{a\})\le 0.3, \qquad 0.2\le P(\{b\})\le 0.9, \qquad 0.1\le P(\{c\})\le 0.5.$$

Линейное представление тела свидетельств

Любое тело свидетельства $F=(\mathcal{A},m)$ можно представить в виде

$$F=\sum_{A\in\mathcal{A}} m(A)F_A,$$

т. е. как выпуклую комбинацию категоричных тел свидетельств.

Например, тело свидетельств из примера можно записать как

$$F=0.2F_{\{b\}}+0.1F_{\{c\}}+0.3F_{\{a,b\}}+0.4F_{\{b,c\}}.$$

Простое тело свидетельств можно представить в виде

$$F_A^{\omega}=(1-\omega)F_A+\omega F_X, \text{ где } \omega\in[0,1].$$

В частности, $F_A^0=F_A$ и $F_A^1=F_X$.

Пигнистическая вероятность

$F=(\mathcal{A},m)$ можно поставить в соответствие **пигнистическую вероятность** $Bet\in\mathcal{P}(X)$, значения которой на $x_i\in X$, $i=1,...,n$ равны вероятности при условии, что в пределах фокальных элементов случайные величины распределены равномерно:

$$Bet(x_i)=\sum_{A\in\mathcal{A},\,x_i\in A}\frac{m(A)}{|A|},\ i=1,...,n.$$

Эти значения совпадают со значениями Шепли

$$Bet(x_i)=\sum_{A\in\mathcal{A},\,x_i\in A}\frac{\left(n-|A|\right)!\left(|A|-1\right)!}{n!}\left(Bel(A)-Bel(A\setminus\{x_i\})\right)$$

соответствующей функции доверия Bel. Пигнистическая вероятность произвольного множества $B\subseteq X$ будет равна

$$Bet(B)=\sum_A\frac{|A\cap B|}{|A|}m(A).$$

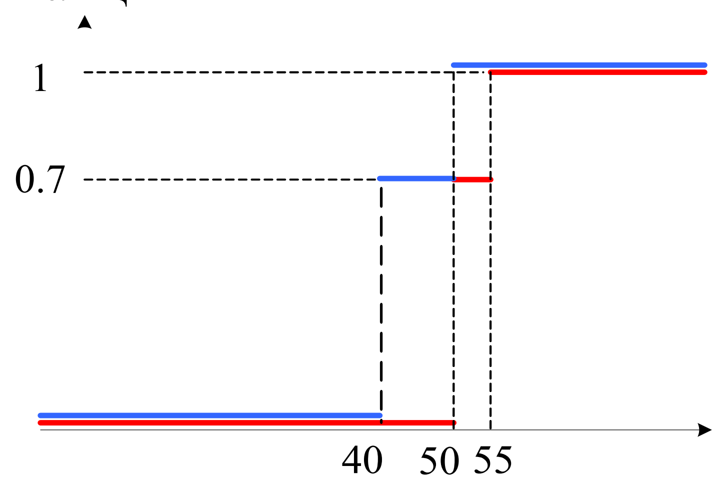
Верхняя и нижняя функции распределения

Если экспертные оценки – ограниченные числовые множества, то можно вычислить верхнюю и нижнюю функции распределения

$$\underline{F}(x) = Bel\{(-\infty, x]\} = \sum_{A \subseteq (-\infty, x]} m(A) = \begin{cases} \sum\limits_{A:\sup A \le x} m(A), & x < \sup X, \\ 1, & x \ge \sup X \end{cases}$$

$$\overline{F}(x) = Pl\{(-\infty, x]\} = \sum_{A \cap (-\infty, x] \ne \varnothing} m(A) = \begin{cases} \sum\limits_{A:\inf A \le x} m(A), & x > \inf X, \\ 0, & x \le \inf X \end{cases}$$

$(\underline{F}(x), \overline{F}(x))$ – *p*-блоки (*p*-boxes). Например, для задачи о прогнозировании стоимости акций



Верхнее и нижнее математические ожидания

Нижняя граница МО: $\underline{\mathbf{E}}[F] = \int_X s d\overline{F}(s) = \sum_A m(A) \inf(A)$.

Верхняя граница МО: $\overline{\mathbf{E}}[F] = \int_X s d\underline{F}(s) = \sum_A m(A) \sup(A)$.

Для задачи о прогнозировании цены акций

$$\underline{\mathbf{E}}[F_1] = 0.7 \cdot 40 + 0.3 \cdot 50 = 43, \qquad \overline{\mathbf{E}}[F_1] = 0.7 \cdot 50 + 0.3 \cdot 55 = 51.5$$

и

$$\underline{\mathbf{E}}[F_2] = 0.6 \cdot 40 + 0.4 \cdot 48 = 43.2, \qquad \overline{\mathbf{E}}[F_2] = 0.6 \cdot 48 + 0.4 \cdot 52 = 49.6.$$

Конфликтность источников информации

Задача. Предположим, что есть свидетельства двух групп экспертов:

$$F_1 = 0.2F_{\{b\}} + 0.1F_{\{c\}} + 0.3F_{\{a,b\}} + 0.4F_{\{b,c\}}, \qquad F_2 = 0.6F_{\{b\}} + 0.4F_{\{a,c\}}.$$

Необходимо оценить конфликтность информации от двух групп экспертов.

Исторически первой (и наиболее популярной) оценкой конфликтности является величина, вычисляемая в правиле Демпстера:

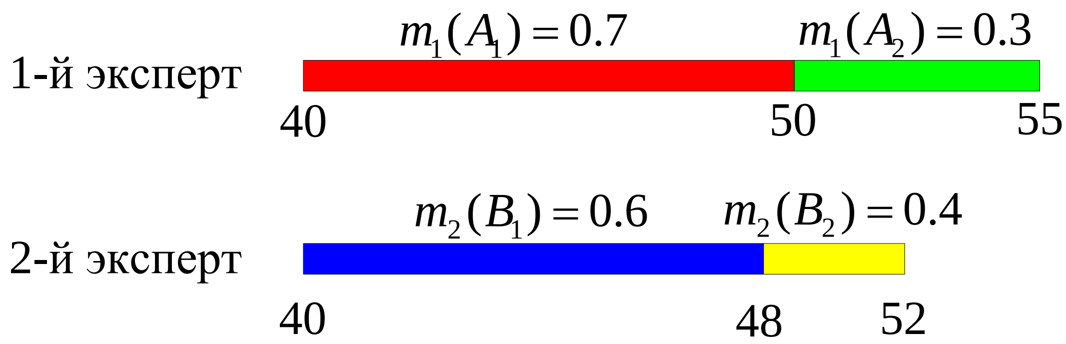
$$K_0(F_1, F_2) = \sum_{A \cap B = \varnothing} m_1(A) m_2(B).$$

$K_0 \in [0,1]$ и $K_0 = 1$ соответствует абсолютной конфликтности свидетельств: $B \cap C = \varnothing$ для всех $B \in \mathcal{A}_1, C \in \mathcal{A}_2$,

В задаче $K_0(F_1, F_2) = 0.2 \cdot 0.4 + 0.1 \cdot 0.6 = 0.14$.

Конфликтность. Примеры

Задача. Два эксперта дали следующую информацию о прогностической стоимости акций



Тогда конфликт между этими двумя свидетельствами равен $K_0 = 0.3 \cdot 0.6 = 0.18$.

Исследование конфликтности свидетельств (аксиоматика, способы задания, внешний и внутренний конфликты, робастные оценки конфликта и т.д.) оформилось в самостоятельное научное направление: A. Martin (2012), S. Destercke & T. Burger (2013), A. Bronevich et al. (2015), W. Liu (2006), A. Lepskiy (2013, 2016, 2017), M. Daniel (2014), J. Schubert (2012) и др.

Агрегирование свидетельств

Если есть два независимых источника информации, заданных свидетельствами $F_1 = (\mathcal{A}_1, m_1)$ и $F_2 = (\mathcal{A}_2, m_2)$ на X, то возникает вопрос, как агрегировать эти данные?

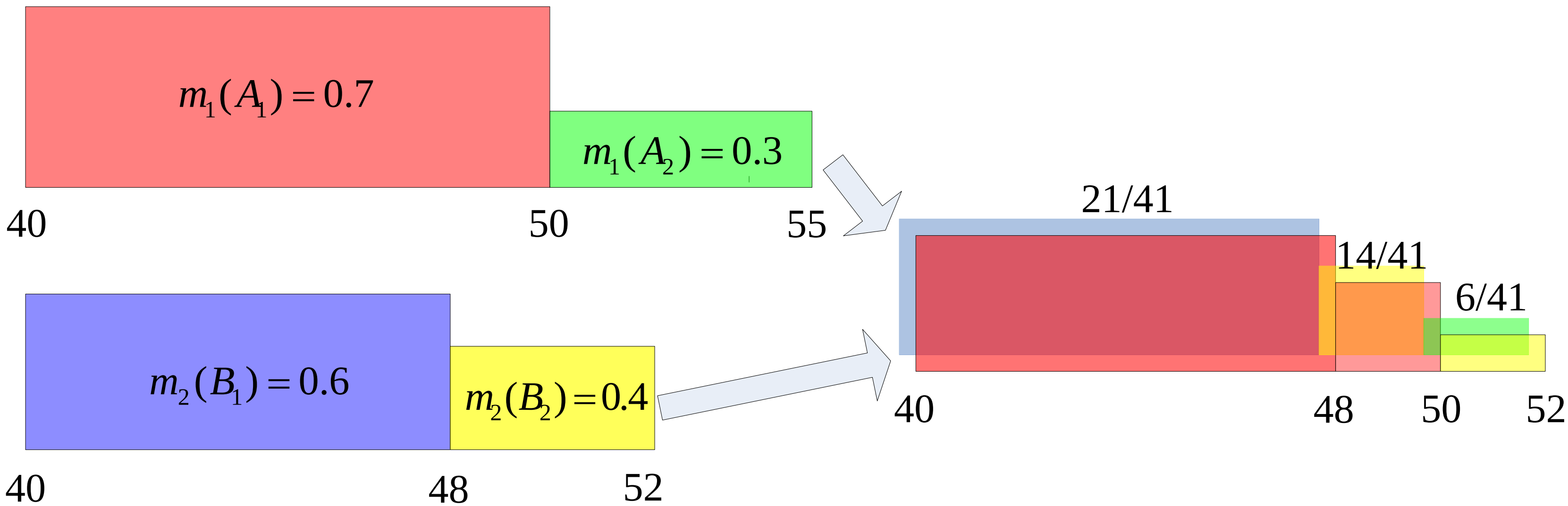
Правило Демпстера (1967). Базовая вероятность $m_D = m_1 \otimes_D m_2$ нового свидетельства вычисляется по формуле

$$m_D(A) = \frac{1}{1-K_0} \sum_{B\cap C=A} m_1(B) m_2(C),\ A \neq \varnothing,$$

где $K_0 = K_0(F_1, F_2) = m_D(\varnothing) = \sum_{B\cap C=\varnothing} m_1\left(B\right) m_2\left(C\right)$. В случае $K_0 = 1$ (абсолютная конфликтность свидетельств) правило Демпстера неприменимо.

Правило Демпстера

Для задачи о прогнозировании стоимости акций



конфликт равен $K_0 = 0.3\cdot 0.6 = 0.18$, а новое свидетельство, полученное с помощью правила Демпстера, будет равно

$m_D([40,48)) = \frac{1}{1-K_0}\cdot 0.7\cdot 0.6 = 21/41$, $\quad m_D([48,50)) = \frac{1}{1-K_0}\cdot 0.7\cdot 0.4 = 14/41$,

$m_D([50,52)) = \frac{1}{1-K_0}\cdot 0.3\cdot 0.4 = 6/41$.

Тогда $\underline{\mathbf{E}} = \frac{21}{41}\cdot 40 + \frac{14}{41}\cdot 48 + \frac{6}{41}\cdot 50 \approx 44{,}19$ и $\bar{\mathbf{E}} = \frac{21}{41}\cdot 48 + \frac{14}{41}\cdot 50 + \frac{6}{41}\cdot 52 \approx 49{,}27$.

Дизъюнктивное правило консенсуса

Правило Демпстера – **оптимистично**: если в соответствии с одним источником информации истинная альтернатива $x \in A$, а в соответствии с другим – $x \in B$, то после применения правила Демпстера – $x \in A \cap B$.

Другим «крайним» случаем агрегирования является **дизъюнктивное правило консенсуса** [Dubois & Prade 1992]

$$m_{\cup}(A) = \sum_{B \cup C = A} m_1(B) m_2(C).$$

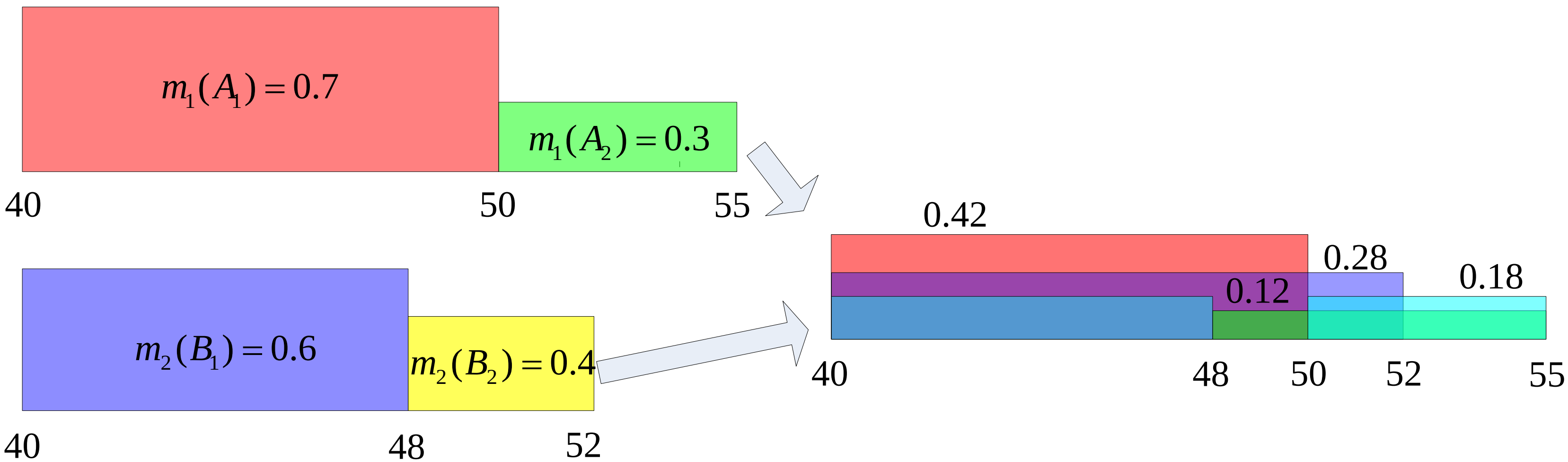
Это правило пессимистично: если один источник информации говорит, что истинная альтернатива $x \in A$, а другой – $x \in B$, то после комбинирования по **дизъюнктивному правилу** – $x \in A \cup B$.

Дизъюнктивное правило консенсуса. Пример

Для задачи о прогнозировании стоимости акций:

$$m_{\cup}([40,50)) = 0.7 \cdot 0.6 = 0.42, \qquad m_{\cup}([40,52)) = 0.7 \cdot 0.4 = 0.28,$$

$$m_{\cup}([40,48)\cup[50,55)) = 0.3 \cdot 0.6 = 0.18, \qquad m_{\cup}([48,55)) = 0.3 \cdot 0.4 = 0.12.$$



Оценки математического ожидания стоимости акции равны:

$$\underline{\mathbf{E}} = 40.96, \ \overline{\mathbf{E}} = 52.6.$$

Оценки являются более осторожными (и более неопределенными), чем в случае комбинирования по правилу Демпстера.

Некоторые другие правила

Правило Ягера [Yager 1987]:

$$m_Y(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \ \ \forall A \neq \varnothing, X \,,$$

$$m_Y(X) = m_1(X) m_2(X) + K_0, \ m_Y(\varnothing) = 0 \,.$$

Конъюнктивные правила Демпстера и Ягера – **оптимистичны**: если в соответствии с одним источником информации истинная альтернатива $x \in A$, а в соответствии с другим – $x \in B$, то после применения правил Демпстера или Ягера – $x \in A \cap B$.

Пример

Предположим, что две независимые группы экспертов дают прогноз о перспективности развития трех технологий $X = \{a, b, c\}$.

Информация первой группы описываются телом свидетельств

$$F_1 = 0.2F_{\{b\}} + 0.1F_{\{c\}} + 0.3F_{\{a,b\}} + 0.4F_{\{b,c\}}.$$

Информация второй группы описывается телом свидетельств

$$F_2 = 0.7F_{\{a,c\}} + 0.3F_{\{a,b,c\}}.$$

Тогда результаты комбинирования будут такими ($\otimes_{ND}$ – ненормализованное правило Демпстера)

	F_1	F_2	$F_1 \otimes_{ND} F_2$	$F_1 \otimes_D F_2$	$F_1 \otimes_Y F_2$	$F_1 \otimes_\cup F_2$
$\varnothing$			0.14			
$\{a\}$			0.21	0.244	0.21	
$\{b\}$	0.2		0.06	0.07	0.06	
$\{c\}$	0.1		0.38	0.442	0.38	
$\{a,b\}$	0.3		0.09	0.104	0.09	
$\{a,c\}$		0.7				0.07
$\{b,c\}$	0.4		0.12	0.14	0.12	
$\{a,b,c\}$		0.3			0.14	0.93
H_0	0.442	0.742	0.132	0.154	0.272	0.974

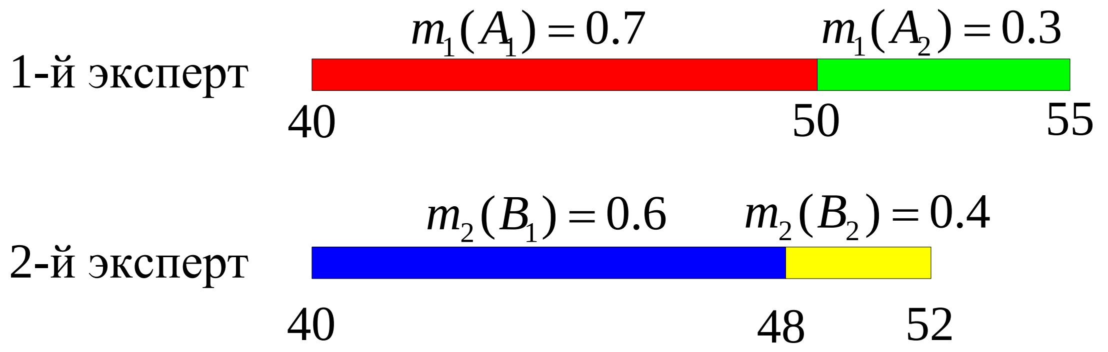
Каноническая мера конфликта для этих свидетельств равна $K = m_{DN}(\varnothing) = 0.14.$

Неопределенность свидетельств

Степень информативной неопределенности оценивают с помощью, так называемых индексов неточности [Bronevich & Lepskiy 2015]. Примером такого индекса является так называемая **обобщенная мера Хартли**:

$$H(\mathbf{m}) = \sum_{A \neq \varnothing} m(A) \ln\left(\left|A\right|\right).$$

Для задачи о прогнозировании стоимости акций двумя экспертами (считая, что $\left|A\right| = l(A)+1$, $l(A)$ – длина интервала)



имеем $H(\mathbf{m}_1) \approx 2.216$ и $H(\mathbf{m}_2) \approx 1.962$, т.е. неопределенность первого свидетельства выше неопределенности второго.

Дисконтирование свидетельств

Предположим, что источники информации имеют разную надежность.

Вопрос: как учесть надежность источников информации?

Правило дисконтирования [Shafer 1976]. Функции масс дисконтируются с коэффициентом $\alpha \in [0,1]$ по правилу:

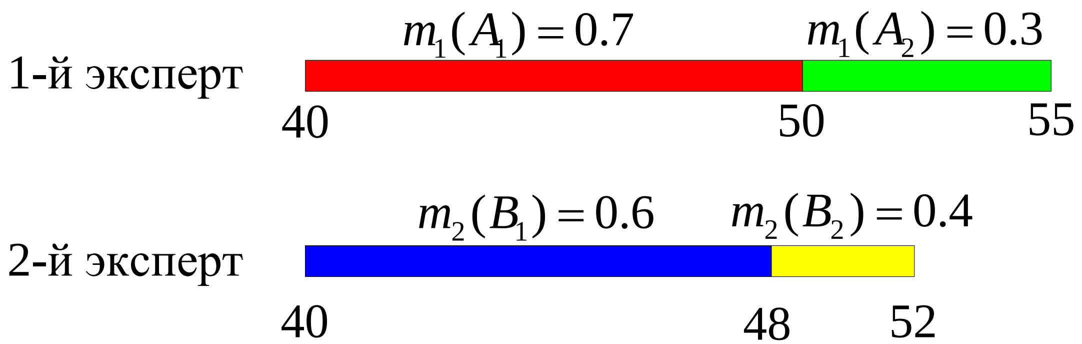
$$m^{\alpha}(A) = (1-\alpha)m(A),\ A \neq X\ ,\ m^{\alpha}(X) = \alpha + (1-\alpha)m(X).$$

Если $\alpha = 0$, то источник информации считается абсолютно надежным. Если же $\alpha = 1$, то источник информации абсолютно ненадежен.

Смысл правила: при уменьшении надежности (увеличении α) функции масс всех собственных подмножеств равномерно уменьшаются, а для несобственного подмножества – увеличивается, что соответствует увеличению информационной неопределенности.

Пример дисконтирования

Задача. Предположим, что в задаче о прогностической стоимости акций



надежность прогноза первого эксперта равна 100% ($\alpha=0$), а второго – 70% ($\alpha=0.3$). Тогда после дисконтирования второго источника, получим:

$$m_2^{0.3}(B_1)=0.7\cdot 0.6=0.42,\ m_2^{0.3}(B_2)=0.7\cdot 0.4=0.28,\ m_2^{0.3}(X)=0.3.$$

После этого конфликтность источников информации уменьшится с $K_0=0.3\cdot 0.6=0.18$ до $K_0^{0.3}=0.3\cdot 0.42=0.126$.

Приложение 1. Принятие решений в торговых системах (совм. с А. Суеваловым)

- $X = \{\text{Sell},\text{Hold},\text{Buy}\}$ – универсальное множество;
- $\mathcal{A} = \{\{\text{Sell}\},\{\text{Sell, Hold}\},\{\text{Hold, Buy}\},\{\text{Buy}\}\}$ – множество фокальных элементов;
- функции масс определяются по правилам Мамдани

$$\text{IF } I_i \text{ IS } T_j \text{ THEN } m_{i,k}(A_s) = \mu_{I_i}^{T_j}(t_k^{(s)}),$$

где $\mu_{I_i}^{T_j}(t_k)$ – функция принадлежности технического индикатора I_i с лингвистической переменной $T_j \in \{$‘Very Low’,‘Low’,‘High’, ‘Very High’$\}$, $t_k^{(s)}$ – k-е значение обучающей выборки для $A_s \in \mathcal{A}$;

- дисконтирование свидетельств

$$m^{\alpha}(A) = (1-\alpha)m(A),\ A \neq X,\ m^{\alpha}(X) = \alpha + (1-\alpha)m(X)$$

(если $\alpha = 1$, то источник абсолютно ненадежный, а если $\alpha = 0$, то источник абсолютно надежный);

Принятие решений в торговых системах

- обучение системы: найти коэффициенты дисконтирования $\{\alpha_i\}$, которые минимизировали бы функционал

$$\sum_k d\left(\mathcal{D}\left(\oplus_{R\,i=1}^{\;n} m_{i,k}^{\alpha_i}\right), \varphi(t_k)\right)^2 \to \min,$$

где $\oplus_R$ – правило комбинирования (н/р, Демпстера и др.); $\mathcal{D}$ – правило принятия решений, $\mathcal{D}: F = (m_R, \mathcal{A}) \mapsto \{\text{Sell, Hold, Buy}\}$, $\varphi(t_k)$ – наилучшее решение из $\{\text{Sell, Hold, Buy}\}$ для обучающего значения t_k, d – метрика.

Результаты

	Without discounting		With discounting		
	Averaging	Yager rule	Averaging	Yager rule	Dempster rule
USD/RUB	86.23	22.00	148.42	84.17	78.97
EUR/USD	21.06	6.71	49.64	58.68	65.76

Приложение 2. Агрегирование классификаторов (совм. с К.Кузнецовым)

Задача. Даны K различных классификаторов для классов $C = \{c_j\}_{j=1}^{M}$, обученных на тестовой выборке $X = \{x\}_{i=1}^{N}$. Требуется агрегировать результаты их работы.

Правило отнесения объекта классу после агрегирования:

$$\hat{c}(x) = \arg max_{c_j \in C} \sum_k w_k \delta\left(\widehat{c_k}(x_i), c_j\right),$$

где $\widehat{c_k}(x_i)$ – предсказание k-го классификатора для объекта x_i, $\hat{c}(x)$ – предсказание ансамбля; $\delta(a, b)$ – символ Кронекера.

Традиционные агрегаторы:

- Plurality vote: $w_k = \frac{1}{K}, \forall k$.
- Simple weighted vote: $w_k = \frac{a_k}{\sum_l a_l}$, где a_k – доля правильно классифицированных объектов k-м классификатором.

Агрегирование классификаторов

- Re-scaled weighted vote. $w_k \propto a_k = \max\left\{0, 1 - \frac{Me_k}{N(M-1)}\right\}$,

  где e_k – количество ошибок допускаемых k-м классификатором.
- Best-worst weighted vote: $w_k \propto a_k = 1 - \frac{e_k - e_B}{e_W - e_B}$,

  где $e_B = min_k\{e_k\}, \quad e_W = max_k\{e_k\}$.
- Quadratic best-worst weighted vote: $w_k \propto a_k = \left(\frac{e_k - e_B}{e_W - e_B}\right)^2$.
- Weighted majority vote: $w_k \propto log\frac{\alpha_k}{1-\alpha_k}$,

  где α_k – точность одиночного классификатора.

Агрегирование классификаторов

Исследуемые классификаторы:

- метод *k* ближайших соседей (knn);
- логистическая регрессия (lr);
- случайный лес (rfc);
- машина опорных векторов (SVM);
- наивный байесовский классификатор (nb).

Построение тела свидетельств для бинарного классификатора:

- $X = \{C_1, C_2\}$ – универсальное множество двух классов;
- $\mathcal{A} = \{\{C_1\}, \{C_2\}, X\}$ – множество фокальных элементов;
- функции масс:

$$m(\{C_1\}) = ap_{C_1}(t), m(\{C_2\}) = ap_{C_2}(t), m(X) = 1 - m(\{C_1\}) - m(\{C_2\}),$$

где a – точность классификации, $p_{C_i}(t)$ – вероятность отнесения образа t классу C_i.

Агрегирование классификаторов

Результаты агрегирования без дисконтирования

Базы Данных/Алгоритмы классификации	Dempster	NB	RFC	LR	SVM	knn	Yager
Pime	**0.837**	0.808	0.808	0.811	0.8105	0.789	0.836
Blood-trans	0.751	0.706	0.679	**0.759**	0.689	0.739	0.75
Banknote auth	**1.00**	0.935	0.991	0.992	0.995	0.995	**1.00**
Spambase	**0.994**	0.856	0.987	0.973	0.972	0.972	**0.994**
Skin segm	**0.999**	0.937	**0.999**	0.908	**0.999**	0.998	**0.999**
Phoneme	0.931	0.824	**0.959**	0.818	0.916	0.914	0.931
Ionosphere	0.935	0.851	0.941	0.839	**0.974**	0.782	0.935
Diabetes	**0.743**	0.728	0.725	0.709	0.726	0.684	0.741
Oil spil	**0.871**	0.604	0.632	0.808	0.860	0.725	**0.871**
Electricity	0.859	0.704	**0.899**	0.734	0.787	0.801	0.859

Приложение 3. Агрегирование рекомендаций (совм. с Е. Кутыниной)

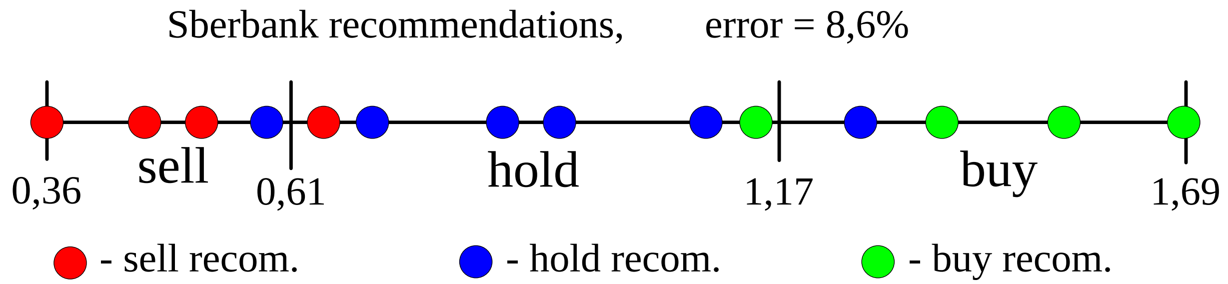
- множество рекомендаций финансовых аналитиков (7 российских банков и 3 аналитические компании) относительно акций российских компаний («голубые фишки») вида {"sell", "hold", "buy"} + целевая цена по акции к концу периода;
- данные по стоимости ценных бумаг за предыдущий период;
- история рекомендаций финансовых аналитиков.

Тело свидетельств j-го аналитика:

$$X = [0, t_{\max}],\ t \in X\text{, где } t = \frac{\text{целевая цена бумаги}}{\text{реальная цена бумаги на момент прогноза}};$$

$F_j = (\mathcal{A}_j, m_j)$, где $m_j(A)$ – относительное число рекомендаций,

$A \in \mathcal{A}_j = \{S_j, H_j, B_j\}$,

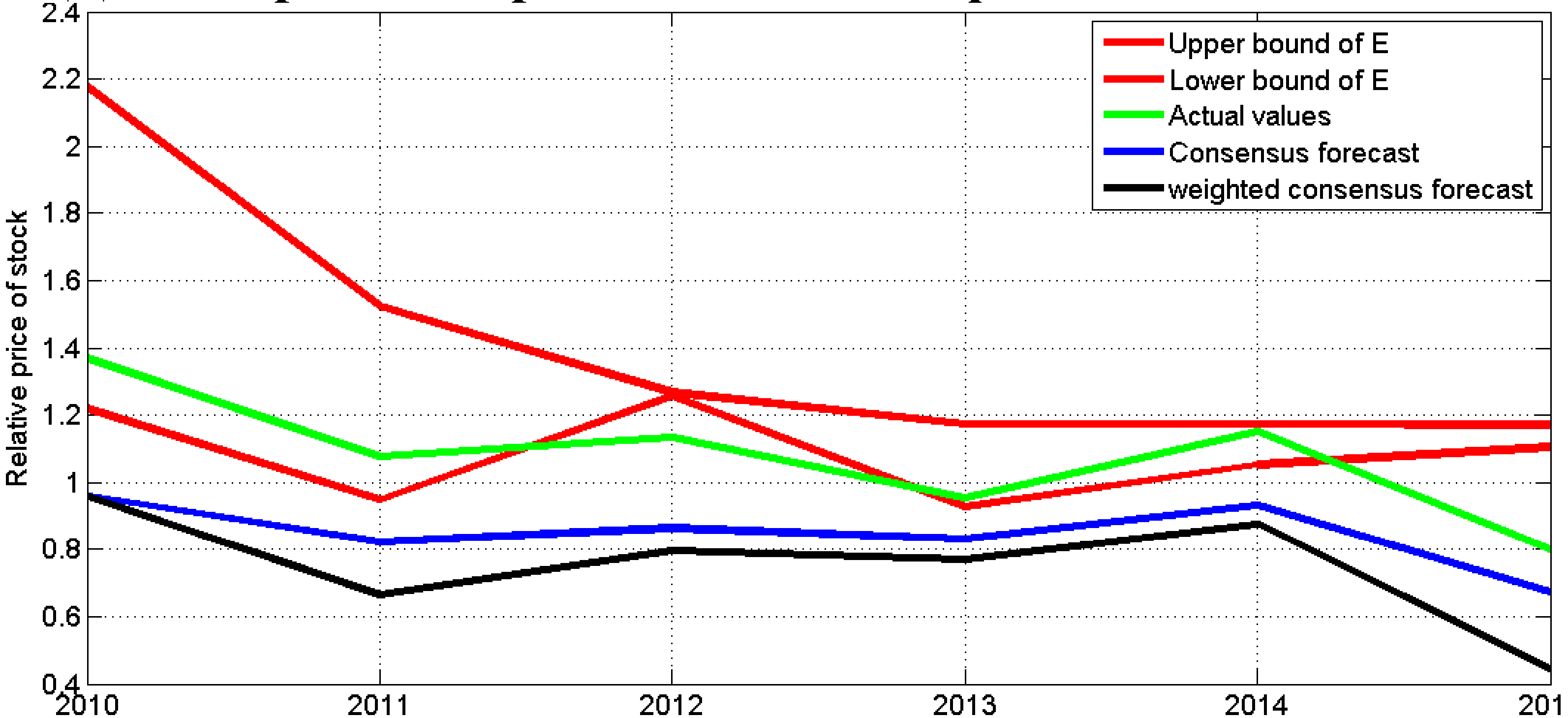


Агрегирование рекомендаций

Стратегии выбора аналитиков для агрегирования

1. Выбираются аналитики с наименьшим конфликтом.
2. Выбираются аналитики с наибольшей надежностью, не конфликтующие между собой.

Прогноз целевой цены акций Транснефти. Агрегирование – правило Демпстера с выбором наименее конфликтных аналитиков



Агрегирование рекомендаций

MAE_{stock}	**WCF**	**CF**	$\underline{E}$, **OSWC**	$\bar{E}$, **OSWC**	E_0, **OSWC**	$\underline{E}$, **NSWC**	$\bar{E}$, **NSWC**	E_0, **NSWC**	$\underline{E}$, **CLCS**	$\bar{E}$, **CLCS**	E_0, **CLCS**
GAZP	0,941	0,456	0,686	0,715	0,653	0,745	0,772	0,711	**0.275**	**0.359**	**0.308**
LKOH	0,499	0,27	**0,19**	0,99	0,59	**0,234**	0,918	0,524	**0.19**3	0.309	**0.263**
ROSN	0,587	0,412	**0,153**	0,486	**0,302**	**0,194**	0,529	**0,265**	**0.110**	**0.257**	**0.191**
SBER	0,55	0,413	**0,137**	0,539	**0,33**	**0,178**	0,633	**0,349**	**0.116**	**0.319**	**0.233**
MAGN	0,298	**0,205**	0,385	0,937	0,661	0,347	1,014	0,681	0.345	0.470	0.422
SNGSP	0,485	0,332	**0,327**	0,84	0,546	**0,243**	1,037	0,598	**0.263**	0.405	**0.330**
GMKN	0,512	0,46	**0,382**	0,673	**0,447**	**0,349**	0,672	**0,345**	**0.283**	**0.314**	**0.267**
VTBR	0,637	0,249	0,451	0,376	0,344	0,525	0,273	0,286	0.269	**0.213**	**0.207**
TRNFP	0,33	0,234	**0,146**	0,356	**0,18**	**0,199**	0,416	**0,18**	**0.119**	**0.215**	**0.141**
TATN	0,568	0,3	**0,198**	0,998	0,589	**0,183**	1,027	0,596	**0.200**	0.421	0.331
MTSS	0,516	0,371	**0,268**	**0,352**	**0,249**	0,39	0,397	**0,259**	**0.214**	**0.205**	**0.210**
CHMF	0,311	0,203	**0,17**	0,369	0,227	**0,196**	0,405	0,222	**0.192**	0.240	0.225
ALRS	0,216	0,14	**0,119**	0,204	**0,116**	0,156	0,308	0,16	**0.115**	0.162	**0.110**
NVTK	0,236	0,395	0,28	0,396	**0,196**	0,28	0,495	**0,149**	0.273	**0.198**	**0.178**
AFLT	0,123	**0,033**	0,477	0,186	0,222	0,552	0,365	0,193	0.376	0.076	0.218
URKA	0,6	0,523	**0,504**	**0,216**	**0,357**	0,654	**0,285**	**0,339**	**0.338**	**0.244**	**0.271**
MAE	0,463	0,312	**0,305**	0,539	0,376	0,339	0,597	0,366	**0.230**	**0.275**	**0.244**

Remark. WCF – Weighted consensus forecast; CF – consensus forecast; the Dempster's rule with discounting and (a) optimistic scenario without censorship (OSWC); (b) neutral scenario without censorship (NSWC); (c) with the choice of the least conflicting sources (CLCS).

Приложение 4. Анализ значимости позиций партий для голосования (совм. с В. Смолевым) Постановка основных задач

В стране ожидаются выборы по партийным спискам в парламент. Перед выборами партии официально отвечают на некоторое множество вопросов, составленное по результатам опросов избирателей.

Задачи:

1) найти множества вопросов, ответы на которые наиболее значимо влияют на результаты голосования;
2) оценить значимость позиций партий по актуальным вопросам на результат голосования за эти партии;
3) оценить конфликтность позиций партий по значимым вопросам;
4) оценить политическую неоднородность общества.

Математическая постановка основных задач

- имеется r партий, которые на выборах набрали $p_1,\dots,p_r$ доли голосов соотв. и $\sum_{i=1}^{r} p_i = 1$;
- есть множество из n вопросов $N_n = \{1,\dots,n\}$, на каждый из которых партия отвечает либо «да», либо «нет»;
- 2^{N_n} – множество всех различных подмножеств вопрос;
- для каждого подмножества вопросов $A \in 2^{N_n}$ надо определить его значимость $m(A)$: $m(A) \geq 0 \ \forall A \in 2^{N_n}$ и $\sum_A m(A) = 1$.

Пусть для $k = 1,\dots,r$, $i = 1,\dots,2n$

$$v_{k,i} = \begin{cases} 1, \text{ если } i \in \{1,\dots,n\} \text{ и } k\text{-я партия ответила "да" на } i\text{-й вопрос}, \\ 1, \text{ если } i \in \{n+1,\dots,2n\} \text{ и } k\text{-я партия ответила "нет" на } (i-n)\text{-й вопрос}, \\ 0, \text{ в противном случае.} \end{cases}$$

Тогда $V_k = (v_{k,1},\dots,v_{k,n},v_{k,n+1},\dots,v_{k,2n})$ – профиль ответов на вопросы k-й партии.

Методология: нахождения наиболее значимых для голосования групп вопросов

Тогда величина

$$C(V_k) = \sum\nolimits_A m(A) \min_{i \in A} \{v_{ki}\}$$

характеризует потенциальное число избирателей k-й партии.
В теории неаддитивных мер такая конструкция называется **интегралом Шоке** [Choquet 1953] от функции $V_k = (v_{k,i})$ по мере $Bel(A) = \sum\nolimits_{B \subseteq A} m(B)$ (мера доверия).

Необходимо найти такие $\{m(A)\}$, чтобы $\dfrac{C(V_k)}{C(V_j)} \approx \dfrac{p_k}{p_j} \;\; \forall k, j \in \{1, \ldots, r\}$.

Примеры критериев минимизации

$K_1 = \max_i \left| p_{i+1} C(V_i) - p_i C(V_{i+1}) \right|$ или $K_2 = \sum_i \left(p_{i+1} C(V_i) - p_i C(V_{i+1}) \right)^2$.

Методология: нахождения наиболее значимых для голосования групп вопросов

Решение задачи минимизации функционала K_2 равносильно решению задачи

$$\begin{cases} \mathbf{m}^T\mathbf{Q}\mathbf{m} \to \min, \\ m_j \geq 0 \;\; \forall j, \\ \sum_j m_j = 1, \end{cases}$$

где $\mathbf{Q} = \sum_{k=1}^{r-1}(p_{k+1}\mathbf{c}_k - p_k\mathbf{c}_{k+1})^T(p_{k+1}\mathbf{c}_k - p_k\mathbf{c}_{k+1})$, $c_{k,j} = \min\limits_{i\in A_j}\{v_{k,i}\}$.

В результате мы получим множество подмножеств вопросов $\mathcal{A} = \{A\}$ (**фокальные элементы**) и их (ненулевые) веса $m(A)$. Пара $F = (\mathcal{A}, m)$ называется **телом свидетельств.**

Методология: нахождения наиболее значимых для голосования отдельных вопросов

Вклад отдельных вопросов в общую значимость подмножеств вопросов можно оценить с помощью **значений Шепли** [Shapley 1953]

$$v_i = \sum_{A \subseteq N_n,\, i \in A} \frac{(n - |A|)!(|A| - 1)!}{n!} \left(Bel(A) - Bel(A \setminus \{i\}) \right).$$

Известно, что для функций доверия вектор Шепли совпадает с так называемым **пигнистическим преобразованием** (pignistic transformation):

$$v_i = \sum_{A \subseteq N_n,\, i \in A} \frac{m(A)}{|A|}.$$

Методология: оценка политической неоднородности общества. Меры конфликта

Мера **внешнего конфликта** между свидетельствами $F_1 = \left(\mathcal{A}_1, m_1\right)$ и $F_2 = \left(\mathcal{A}_2, m_2\right)$

$$K = K\left(F_1, F_2\right) = \sum_{A \cap B = \varnothing} m_1(A) m_2(B).$$

$K \in [0,1]$ и значение $K = 1$ соответствует абсолютной конфликтности двух тел свидетельств: $A \cap B = \varnothing$ для любых $A \in \mathcal{A}_1$ и $B \in \mathcal{A}_2$.

Меры **внутреннего конфликта** свидетельства $F = \left(\mathcal{A}, m\right)$:

1) автоконфликт $K_{aut}(F) = K(F, F)$

2) $K_{\text{int}}(F) = \min\limits_i Bel(N_n \setminus \{i\})$.

Методология: оценка влияния позиций партий на голосование

Подмножество $A \in 2^{\overline{N_n}}$ содержится в профиле k-й партии V_k (обозн. $A \subseteq V_k$), если $\forall i \in A : v_{k,i} = 1$.

Пусть $\mathcal{A}_k = \{A \in \mathcal{A} : A \subseteq V_k\}$ – проекция фокальных элементов на профиль k-й партии и $\overline{m}_k(A) = \overline{m}(A)$, если $A \in \mathcal{A}_k$;

- $\overline{m}_k(A)$ характеризует электоральную значимость для k-й партии позиции по множеству вопросов $A \in \mathcal{A}_k$.
- $M_k = \sum_{A \in \mathcal{A}_k} \overline{m}_k(A)$ характеризует степень значимости позиции k-й партии на результат голосования за нее;
- $M_k \in [0,1]$; если $M_k = 0$, то позиция партии по списку вопросов незначима для выборов; если $M_k = 1$, то позиция партии по всем вопросам значима для победы на выборах.

Методология: оценки конфликтности позиций партий по значимым вопросам

По функциям множеств $\{\bar{m}_k\}$ сформируем свидетельства $\{F_k\}$, $F_k = (\mathcal{A}_k, m'_k)$, $m'_k = \bar{m}_k / M_k$, если $M_k \neq 0$. Если же $M_k = 0$, то это говорит о незначимости позиции партии на результат выборов.

Под конфликтностью (несогласованностью) позиций между i-й и j-й партиями будем понимать величину $K_{i,j} = K(F_i, F_j)$, $K_{i,j} \in [0,1]$.

Значение $K_{i,j} = 1$ означает, что позиции i-й и j-й партий по всем вопросам и их различным значимым для голосования комбинациям не пересекаются.

А значение $K_{i,j} = 0$ означает, что у i-й и j-й партий нет значимых для голосования непересекающихся позиций.

Описание данных: выборы в Бундестаг ФРГ в 2013г.

Рассмотрены 6 партий, набравших более 3% голосов:

- блок Христианско-демократического союза и Христианско-социального (CDU/CSU),
- Социал-демократическая партия Германии (SPD),
- «Левые» (DIE LINKE),
- «Союз 90/Зелёные» (GRUNE)
- Свободная демократическая партия (FDP),
- «Альтернатива для Германии» (AfD).

Описание данных: выборы в Бундестаг ФРГ в 2013г.

Рассмотрены 8 самых «весомых» вопросов из 38 (сервис опроса граждан http://www.bpb.de/politik/). Примеры вопросов:

- #22 о повышении пенсионного возраста в стране (3 490 000 голосов);
- #31 об отношении к требованию для всех граждан обязательно регистрироваться в государственной системе медицинского страхования (3 160 000 голосов);
- #18 – должен ли быть уровень государственного финансирования студентов независим от достатка родителей (2 060 000 голосов);
- #1 – вводить ли общенациональную минимальную оплату труда (2 020 000 голосов);
- #32 – должны ли государства в Еврозоне оплачивать самостоятельно свои долги (1 450 000 голосов);
- #20 – нужно ли введение обязательной квоты для женщин в совете директоров компаний (1 220 000 голосов).

Результаты: наиболее значимые группы вопросов

- для множества из 8-ми вопросов и 6 партий (или блоков) было выявлено 1003 подмножеств вопросов A, для которых $\min_k \min_{i \in A} \{v_{k,i}\} > 0$ (реальная размерность векторов $\mathbf{m}$);
- найдено 255 подмножеств вопросов с ненулевыми значениями функции масс;
- группа наиболее значимых подмножеств состоит из 30 наборов вопросов со значимостью 0.0072;
- во все подмножества с наибольшей значимостью вошли вопросы #18 (должен ли быть уровень государственного финансирования студентов независим от достатка родителей), #22 (о повышении пенсионного возраста) и #2 (о субсидировании семей с детьми, не посещающих учебные заведения);
- значения Шепли будут равны 0.18, 0.15, 0.12 для вопросов #2, #8, #22 соответственно и примерно 0.11 для остальных 5-ти вопросов.

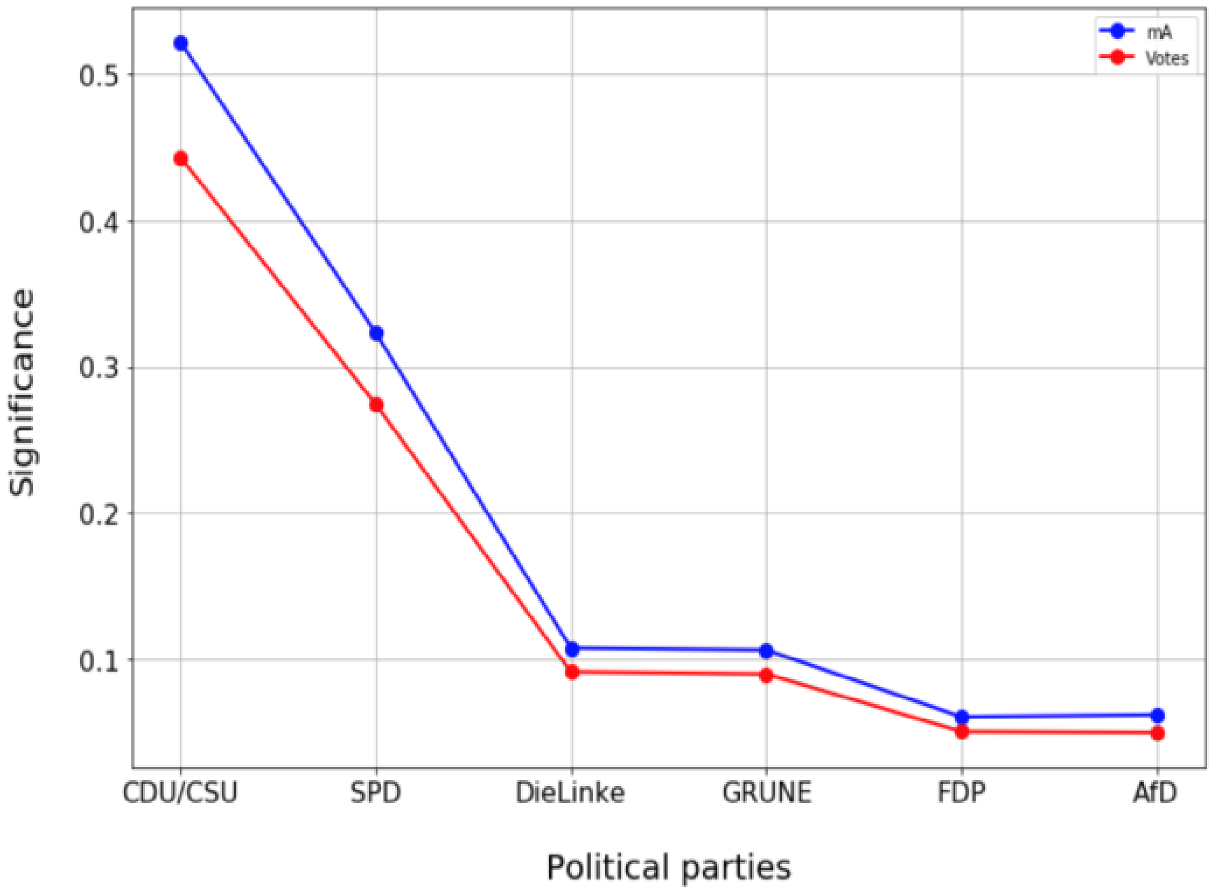
Результаты: оценивание политической неоднородности общества

Мера внутреннего конфликта свидетельства значимости вопросов. равна

$$K_{\mathrm{int}}(F) = 0.264.$$

Это говорит об определенной политической однородности общества.

Результаты: оценка значимости позиций отдельных партий для голосования



Результаты: оценка конфликтности позиций партий по значимым вопросам

чем «краснее» – тем конфликтнее

	CDU/CSU	SPD	DieLinke	GRUNE	FDP	AfD
CDU/CSU		0,35	0,59	0,42	0,14	0,14
SPD			0,12	0,16	0,40	0,20
DieLinke				0,13	0,32	0,48
GRUNE					0,20	0,36
FDP						0,18
AfD						

Список литературы

1. Bronevich A., Lepskiy A., Penikas H. The Application of Conflict Measure to Estimating Incoherence of Analyst's Forecasts about the Cost of Shares of Russian Companies// Procedia Computer Science, 2015, 55, pp.1113-1122.
2. Dempster A.P. Upper and Lower Probabilities Induced by a Multivalued Mapping// The Ann. of Math. Stat., 1967, 38(2), 325–339.
3. Dubois D., Prade H. On the combination of evidence in various mathematical frameworks// In: Flamm, J., Luisi, T. (eds.) Reliability Data Collection and Analysis, Kluwer Acad. Publ., Dordrecht, 1992, 213–241.
4. Shafer G. A Mathematical Theory of Evidence. – Princeton, N.J.: Princeton University Press, 1976.
5. Lepskiy A., Suevalov A. Application of the Belief Function Theory to the Development of Trading Strategies// Procedia Computer Science, 2019, 162, 235–242.
6. Kutynina E., Lepskiy A. Aggregation of Forecasts and Recommendations of Financial Analysts in the Framework of Evidence Theory// in: Kacprzyk J. et al. (eds). Advances in Intelligent Systems and Computing, 2018, 642, Springer, Cham, 370–381.
7. Yager R. On the Dempster–Shafer framework and new combinations rules// Inform. Sc., 1987, 41, 93–137

Спасибо за внимание!