Определение параметров межзвездного поглощения света по данным каталога Hipparcos

Ф.Амосов

18 апреля 2015 г.

Аннотация

Основная задача исследования — построение метода автоматического поиска пылевых облаков в окрестности Солнца на основе массовых каталогов звезд. На первом этапе метод был применен к данным каталога Hipparcos, а именно для построения двумерной панорамы распределения пылевых облаков на небесной сфере. Метод исследования основан на сравнении эталонного показателя цвета звезды данного спектрального класса с наблюдаемым показателем цвета. Так как полная двумерная спектральная классификация известна не для всех звезд каталога Hipparcos, была решена вспомогательная задача: используя параллаксы звезд каталога Hipparcos дополнить информацию о звездах классом светимости на основе данных о видимой звездной величине и параллаксе звезды. В результате работы была построена карта распределения пылевой материи, вызывающей покраснения света звезд, на небесной сфере. Недостаточная плотность звезд и низкие относительные точности параллаксов для далеких звезд не позволяют полностью использовать возможности метода, которые будут полностью раскрыты по завершении миссии GAIA

1 Введение

Общая задача поиска облаков межзвездной пыли, ответственной за поглощение света звезд, является трехмерной. Для ее надежного выполнения необходим массовый каталог параллаксов звезд. На сегодняшний день таким каталогом является Hipparcos, но, как мы увидим, его точности недостаточно для решения этой задачи. Видимо её полное решение будет возможно только после получения результатов миссии GAIA, чьей основной задачей является определение структуры Млечного Пути в окрестности Солнца. Поэтому мы представим решение частной задачи – задачи построения двумерного распределения пыли по небесной сфере. Ее решение позволит по крайней мере определить направления наибольшего поглощения и изменения таких характеристик звезд как звездная величина и показатель цвета Это позволит в будущем учитывать межзвездное поглощение при определении характеристик звезды таких как, к примеру, показатель цвета.

2 Исходные данные

2.1 Общие сведения о каталоге Hipparcos

В 1989 году Европейское Космическое Агентство (ESA) осуществило запуск космического аппарата HIPPARCOS (HIgh Precision PARallax COllecting Satellite — «спутник для сбора высокоточных параллаксов») с целью получения положений, собственных движений и параллаксов звезд на миллисекундном уровне точности. Космический аппарат проработал на орбите 37 месяцев, в течение которых он выполнял астрометрические и фотометрические измерения звезд по заданной программе. Обработка этих наблюдений привела к созданию двух каталогов: Hipparcos[1], содержащего информацию о 118218 звездах с точностью определения положений, годичных собственных движений и параллаксов на уровне 1 mas (milli arc second), и каталога Tycho[5], содержащего уже свыше 1 млн. звезд, с точностью измерения тех же параметров до 25 mas.

Положения и собственные движения звезд в Hipparcos приводятся в фундаментальной системе ICRS (International Celestial Reference System), реализованной в настоящее время с помощью каталога внегалактических радиоисточников, получившего название ICRF (International Celestial Reference Frame). Следует отметить, что достигнутая точность привязки осей координат системы отсчета каталога HIPPARCOS к осям ICRF оценивается величиной 0.6 mas по всем трем углам поворота и величиной 0.25 mas/год по всем трем компонентам вектора остаточного взаимного вращения двух систем отсчета.

В 2007 году вышла новая редакция астрометрических данных каталога Hipparcos [4] — каталог HIPNEWCAT (HIPparcos NEW astrometric CATalog). Утверждается, что точность положений, параллаксов и собственных движений всех звезд, ярче $H_P = 8$, улучшена в 4 раза, а для всех остальных звезд более, чем в 2 раза. Уменьшена взаимная корреляция параметров иногда в 10 раз. Именно эта версия использовалась в работе в качестве источника астрометрических данных.

2.2 Фотометрические системы каталогов Tycho и Hipparcos

Фотометрические измерения на основном инструменте спутника HIPPARCOS выполнялись в широкой полосе (обозначаемую как H_P). В дополнение, почти для всех звезд каталога была выполнена двухцветная фотометрия (фотометрия Tycho величины V_T и B_T). Точность определения H_P составляет $0.0004^m - 0.007^m$ (для звезд $2 - 12^m$), а точность одного измерения — $0.003^m - 0.05^m$.

Фотометрические системы H_P, V_T и B_T — это инструментальные системы, и они не совпадают с общепринятой системой Джонсона. Используя значения звездной величины V_J по шкале Джонсона и показателя цвета для 8000 стандартных звезд с хорошими фотометрическими данными в системе B_T и V_T, были получены следующие эмпирические линейные соотношения, применимые к диапазон $-0.2 < (B-V)_T < 1.8$:

$$V_J = V_T - 0.090(B-V)_T$$

$$(B-V)_T = 0.850(B-V)_T$$

Точность этих преобразований в среднем лучше, чем 0.015^m для V_J и 0.05^m для $(B-V)_T$. Эти преобразования применимы к звездам, чей цвет не искажен межзвездным поглощением, и игнорируют зависимость от класса светимости. Формулы вообще не применимы к звездам класса M, даже если их показатель цвета $(B-V)_T < 1.8^m$.

2.3 Распределение звезд по абсолютной звездной величине

Измерения блеска звезд в двух полосах B_T, V_T были сделаны для 114820 звезд в Hipparcos, кроме этого для 2117 звезд фотометрические данные были взяты из наземных источников. Для 115180 звезд имеются данные и о спектральном классе. Индивидуальные параллаксы позволяют вычислить абсолютную звездную величину или светимость для каждой звезды.

2.4 Диаграмма Герцшпрунга-Рессела

Каталог Hipparcos дает уникальную возможность построить диаграмму для любой выборки звезд. До появления этого каталога это было невозможно из-за плохого знания расстояний. Диаграммы удавалось строить только для тех звезд, для которых имелась косвенная информация, что они находятся от нас примерно на одинаковом расстоянии, например, для звезд одного звездного скопления. Рис. 1 показывает диаграмму Герцшпрунга-Рессела для звезд, ближе 100 пк, у которых относительные ошибки определения параллаксов не будут превышать 10%.

2.5 Спектральные характеристики звезд

Каталог Hipparcos для большинства звезд содержит информацию о спектральном типе, полученную из наземных наблюдений. Основной источник – Мичиганский каталог (Michigan catalogue for the HD stars, vol. 1-4, (Houk+, 1975-1988)) и несколько других источников.

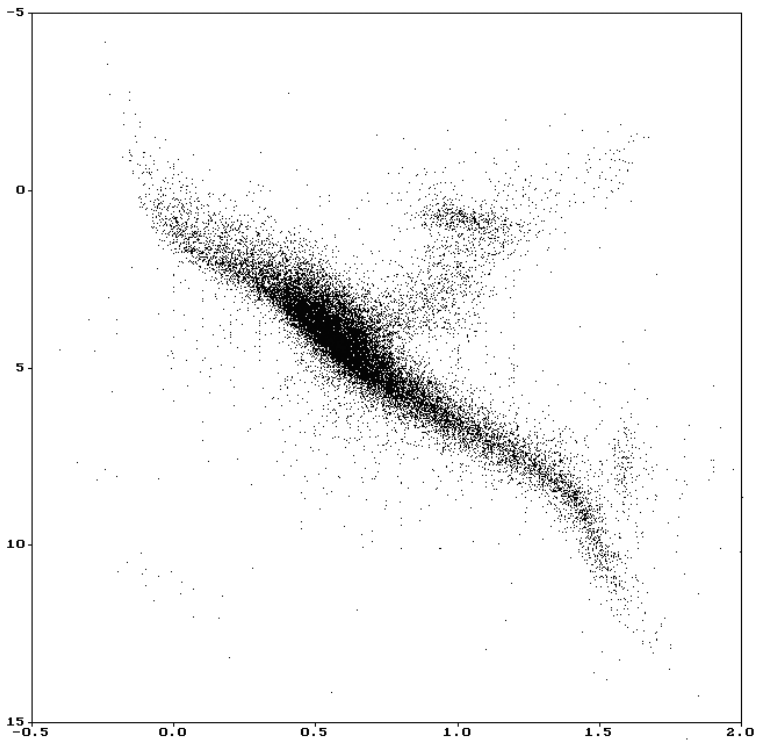

Рис. 1: Диаграмма Герцшпрунга-Рессела для звезд ближе 100 пк на основе данных каталога Hipparcos

Однако, информация о спектральных классах приведена только для звезд южного экваториального полушария (2).

2.6 Используемые данные

Для определения параметров межзвездного поглощения, необходимо иметь следующие данные о звездах,

- положение
- параллакс
- фотометрия
- спектральный класс и класс светимости

В каталоге Hipnewcat отсутствуют спектральные данные, поэтому в данной работе они были взяты из каталога Hipparcos. Так же из Hipparcos были получены значения видимой звездной величины V_{mag}. Итак, мы рассматривали только те звезды, которые имеют данные о положении, параллаксе, покраснении в Hipnewcat и данные о спектральном типе и видимой звездной величине в Hipparcos. Кроме того, мы не рассматривали звезды, у которых в каталоге Hipnewcat указано число компонент более одной. Число звезд, используемых в данной работе получилось равным 98827.

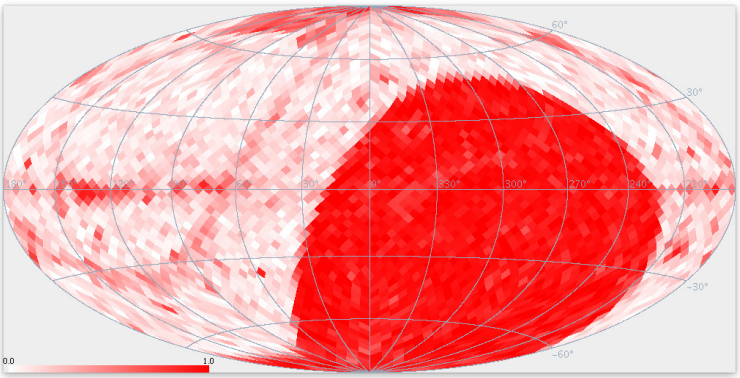

Рис. 2: Распределение наличия класса светимости у звезд каталога Hipparcos на небесной сфере (см. Приложение). Чем больше в пикселе доля звезд, имеющих класс светимости, тем он краснее. Видим отсутствие класса светимости практически у всех звезд северного экваториального полушария и присутствие у южного

3 Покраснение

3.1 Определение

Межзвездное поглощение может быть описано избытком цвета. Избыток цвета мы будем называть «покраснением». *Покраснение* звезды есть

$$E = E_{B-V} = (B-V)_{obs} - (B-V)_{int}, \tag{1}$$

где $(B-V)_{obs}$ — ее видимый нами показатель цвета, а $(B-V)_{int}$ — теоретический показатель цвета звезды. Значение $(B-V)_{obs}$ мы можем получить на основе данных фотометрии звезды из каталога. Величины $(B-V)_{int}$ оцениваются статистически. В нашей работе мы воспользуемся приведенной в [7] двумерной таблицей «спектральный класс, класс светимости — показатель цвета». То есть, для получения $(B-V)_{int}$ звезды нам потребуются ее спектральный класс и класс светимости. Их мы можем получить из данных каталога. Приведем пример расчета покраснения на звезде HIP 44800,

- У нее в каталоге $(B-V)_{obs} = 0.535^m$
- Класс F7V, поэтому (по [7]) $(B-V)_{int} = 0.493^m$
- Покраснение $0.535^m - 0.493^m = 0.042^m$

Покраснение — это количественное измерение межзвездного поглощения, поэтому мы можем сказать, что «между нами и звездой HIP 44800 пыли на 0.042^m».

4 Способ получения классов светимости

Как мы увидели в обзоре данных каталога Hipparcos, практически у всех звезд северного экваториального полушария отсутствует класс светимости. Для нас его наличие чрезвычайно важно, ввиду того, что мы на основе класса светимости и спектрального класса рассчитываем истинное значение $B-V$ для звезд ($(B-V)_{int}$). Тем самым, отсутствие класса светимости у половины звезд делает невозможным проведение наших расчетов для всего северного экваториального полушария.

Исправим это. Определим неизвестные классы светимости. Сделаем это с помощью методов машинного обучения. Натренируем классификатор, который будет выдавать класс светимости для звезды по

двум факторам - ее показателю цвета и ее абсолютной звездной величине. Этих факторов должно быть достаточно, т.к. классы светимости теоретически разделимы на диаграмме Герцшпрунга-Рессела.

Доля звезд, которые относятся ни к III, ни к V классам мала (16,3%, 8058 из 49285, имеющих класс светимости). Поэтому, мы упростим задачу — обучим линейный бинарный классификатор, который будет предсказывать III или V класс. Сделаем это с помощью метода опорных векторов [8] (см. Приложение).

В качестве обучающего множества возьмем все звезды, у которых присутствует класс светимости, и он либо III, либо V. Таких звезд 39807. Изобразим их распределение по классам в виде таблицы

класс светимости	$B-V<0.6$	$B-V\geq 0.6$	всего
III	1947	16681	18628
V	15549	5630	21179
III & V	17496	22311	39807

Обучим на них классификатор — получим разделяющую классы прямую (3).

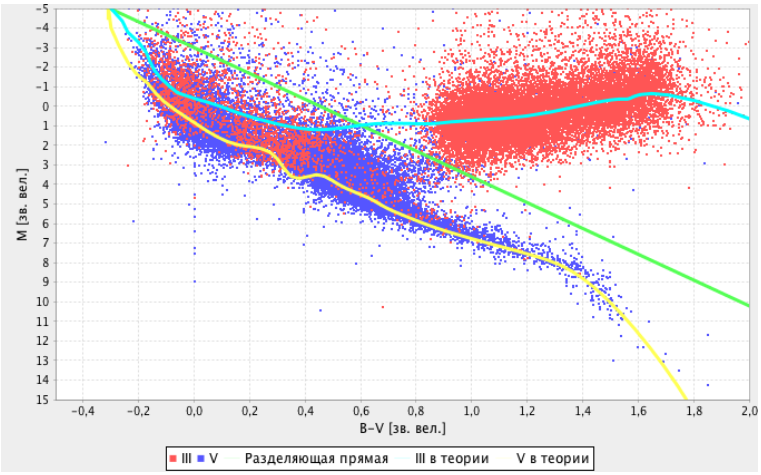

Рис. 3: Обучающее множество с полученной разделяющей прямой. Также здесь изображены теоретические кривые для классов на основе таблиц [7]

Уравнение разделяющей прямой,

$$F(B-V,M)=-2.9876\cdot(B-V)+0.4526\cdot M+1.3547=0$$

где $B-V$ – показатель цвета, M – абсолютная звездная величина.

Мы видим, что при показателе цвета $B-V\geq 0.6$ классификатор работает практически идеально, но при $B-V<0.6$ и $F(B-V,M)>0$ он всем звездам предсказывает V класс при большой доле звезд III класса в обучающем множестве (10.3%, 1758 из 17029). При $B-V<0.6$ звезды III и V класса неразделимы, поэтому для этой половины можно принять другое решение — результат работы классификатора будет взвешенным средним III и V классов, то есть некоторым средним классом, $(B-V)_{int}$, у которого будет равен

$$(B-V)_{int}=w_1\times(B-V)_{int}(III)+w_2\times(B-V)_{int}(V)$$

.

В этой формуле веса w_1 и w_2, $(w_1+w_2=1)$ логично взять в соответствии с априорной вероятностью классов в этой области (10.3%/89.7%).

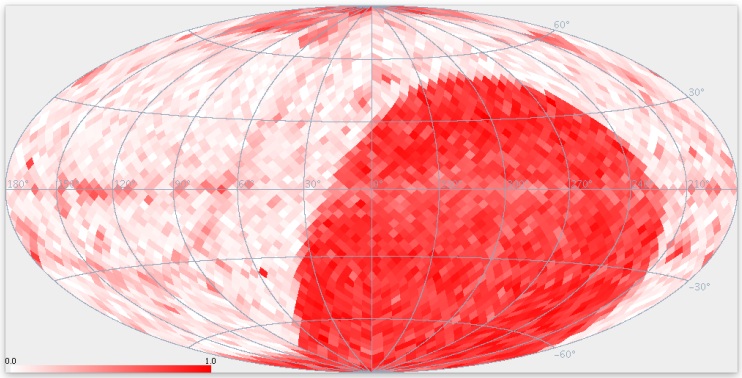

Рис. 4: Распределение обучающего множества по небесной сфере. Отличие от (2) только в том, что здесь звезды только III и V классов

В таблице двумерной спектральной классификации [7] у спектрального типа G2 показатель цвета при III классе 0.733, при V — 0.630. Это максимальная разница между III и V классом в интересующем нас диапазоне. Класс, соответствующий средневзвешенному решению будет иметь показатель цвета

$$(B-V)_{int} = 10.3\% \cdot 0.733 + 89.7\% \cdot 0.630 = 0.640$$

.

Как мы видим, даже в наихудшем случае отличие от V класса минимальное — 0.01^m — гораздо ниже уровня каталожных ошибок. Поэтому в дальнейшем мы всегда будем использовать решение классификатора в этой области по V классу.

Для количественной оценки качества работы классификатора, проведем 10-fold кросс-валидацию.

Ниже приведены ее результаты,

Решение классификатора →	III	V
III	16636	1992
V	783	20396

Класс	Точность	Полнота	F1-мера
III	95%	89%	92%
V	91%	96%	93%

Классификатор имеет приемлемое качество. Результат его работы — наличие класса светимости у всех рассматриваемых звезд.

5 Градиент покраснения по расстоянию

5.1 Идеальная кривая покраснения

Предположим, что на некотором луче зрения бесконечно много звезд, и они расположены на нем всюду плотно. Пусть для каждой звезды мы можем идеально измерить ее покраснение. Тогда, ход покраснения на этом луче зрения должен иметь следующий вид,

Здесь по оси x отложено расстояние по лучу зрения, а по y — покраснение у соответствующих звезд. Покраснение должно всегда монотонно расти, т.к. пыль присутствует всюду. Очевидно, что там, где покраснение растет быстрее - пыли больше, там где медленнее - меньше. Поэтому, можно было бы сказать, что облака на этом луче зрения находятся там, где покраснение растет «очень быстро» (синие области),

Тем самым, построение кривых покраснения в разных направлениях на небе может позволить находить области повышенного мезвездного поглощения, то есть находить пылевые облака.

Рис. 5: Идеальный ход покраснения на луче зрения

Существуют таблицы «спектральный класс, класс светимости — показатель цвета», поэтому для каждой звезды из каталога Hipparcos, у которой есть видимый показатель цвета и спектральный класс, можно вычислить покраснение. Если мы для всех таких звезд знаем еще и их пространственные координаты с хорошей точностью, то мы можем говорить о пространственном распределении покраснения. В данной работе мы рассмотрим это распределение с точки зрения трендов покраснения в различных направлениях. Результатом работы будут коэффициенты a и b этих трендов $ar + b$.

Построение кривых покраснения в разных направлениях позволит нам понять пространственное распределение межзвездного поглощения (пыли). Звезд в каталоге не бесконечное число, поэтому реальные кривые покраснения будут не непрерывными кривыми, а будут наборами точек, описывающими ход покраснения. Аналогично, вместо звезд на луче зрения мы должны использовать звезды в малых конусах. Поэтому, ход покраснения у нас будет выглядеть, к примеру, так,

Следующий метод позволит построить «кривые» покраснения во всех направлениях на небесной сфере. Он состоит из трех этапов,

5.2 Картирование небесной сферы

Обозначим площадки, соответствующие «пикселям» разбиения HEALPix за $\{P_i\}_{i=1}^{N_{side}}$ (у нас $N_{side} = 18$). Обозначим конусы, высекаемые соответствующими пикселями через $\{C_i\}_{i=1}^{N_{side}}$

Такое разбиение позволит нам,

1. Рассмотреть ход покраснения в каждом конусе как одномерную функцию $E(r)$. Это корректно, ввиду того, что конусы достаточно узкие;
2. Сделать наши результаты «независимыми», т.к. конусы не пересекаются;
3. Поместить в каждый конус примерно одинаковое число звезд, чтобы избежать недостатка звезд в некоторых конусах.

5.3 Тренд

Тем самым, мы ищем $E_i(r)$, соответствующую каждому C_i. Ввиду того, что практически все $E_i(r)$ очень сильно зашумлены разного рода ошибками, не удается проследить истинный ход этих функций. Но тренд вида kr все же можно вычислить ($E_i(r) \approx k_i r$). Он находится с помощью метода наименьших квадратов.

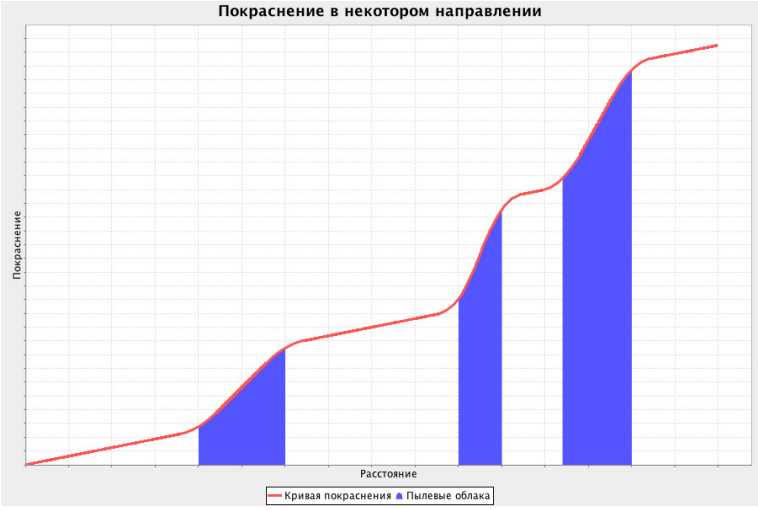

Рис. 6: Облака на идеальной кривой покраснения

5.4 Критерии выбора звезд

5.4.1 По параллаксу

В каталоге Hipparcos, далекие звезды имеют очень большие ошибки в параллаксе. На расстояниях, скажем, в 400 пк они могут достигать 100%. Такие ошибки могут очень сильно испортить наши результаты. Тем самым, мы не будем рассматривать звезды, у которых относительная ошибка параллакса не превосходит 25%. Меньшее значение порога оставит нам очень малое число звезд, которых не хватит для того, чтобы вычисленные тренды были достоверными.

5.4.2 По тренду

Покраснения некоторых звезд вносят большие ошибки в искомые тренды. Иногда они получаются вообще отрицательными, что противоречит здравому смыслу. Но об этом мы поговорим позже.

Как известно, метод наименьших квадратов не устойчив к выбросам, т.е. совсем неверные покраснение/параллакс могут очень сильно испортить тренд. Для более устойчивого построения тренда, мы сделаем следующее. После построения тренда по всем звездам в конусе, мы выбрасываем те, у которых отклонение от тренда самое большое. Затем, мы строим тренд заново, но уже только по оставшимся звездам. После выброса 10% самых плохих звезд тренд, к примеру, может быть таким

В данном случае, мы выбросили одну звезду (синюю). 10% звезд — это обычно 0, 1 или 2 звезды в каждом конусе. Опять же, нельзя выбрасывать слишком много звезд из-за опасности «подгонки» данных под модель.

5.5 Распределение коэффициента k по небесной сфере

Тем самым, ход покраснения в конусе C_i мы описываем одним числом k_i — скоростью роста покраснения в этом конусе. Она, как мы ранее выясняли, должна коррелировать с наличием пыли. Поэтому, составив карту распределения коэффициента k, мы составим панораму пыли в окрестности Солнца.

Это небо в галактической системе координат с центром в центре галактики. Сфера разбита на $12 \cdot 18^2 = 3888$ «пикселей» алгоритмом HEALPix. В каждом кусочке построен тренд покраснения kr. На этом рисунке изображено распределение значения коэффициента k по пикселям. Синий цвет означает отрицательное значение k, красный - положительное. Чем насыщеннее цвет, тем больше значение коэффициента по модулю.

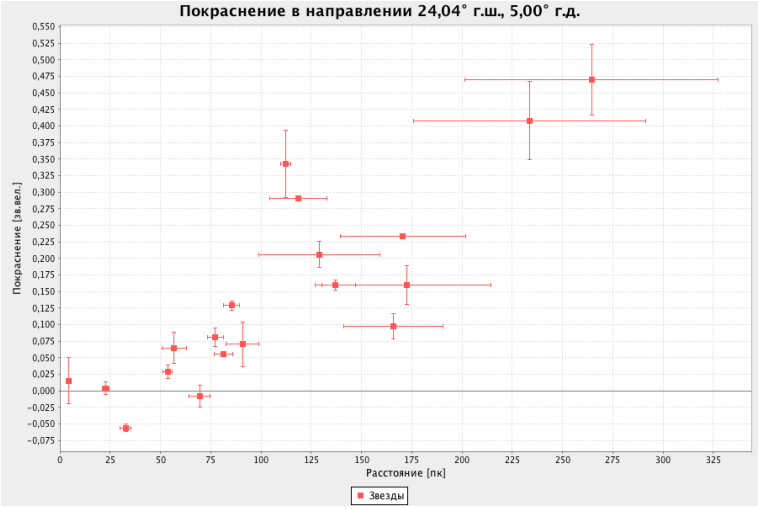

Рис. 7: Покраснение из данных каталога Hipparcos

6 Отрицательное покраснение

На некоторых площадках коэффициент k отрицательный. В теории, такого не должно быть, т.к. межзвездное поглощение не может делать звезды более голубыми. Пример хода покраснения на одной из площадок, Мы видим, что это вызвано тем, что некоторые звезды имеют сильно отрицательное покраснение. То есть $E_{B-V} + 3\sigma_{E_{B-V}} < 0$, где σ — это ошибка. Такого не должно быть, т.к. звезда не может сильно голубеть на расстоянии. Давайте рассмотрим какую-нибудь звезду, имеющую такое аномальное отрицательное покраснение. К примеру, HIP 66713. Ее параметры:

- Параллакс 7.99 ± 0.77
- Показатель цвета 0.386 ± 0.014
- Спектральный тип G0V
- Видимая звездная величина 8.37

Мы видим, что она не имеет аномальных параметров. Так же, ее параметры имеют небольшие ошибки. Согласно таблице в [7], спектральному типу G0V соответствует показатель цвета 0.580. Тем самым, покраснение $E = (B-V)_{obs} - (B-V)_{int} = (0.386 \pm 0.014) - 0.580 = -0,194 \pm 0.014$. Мы видим аномальное покраснение у не аномальной звезды. И таких звезд (имеющих $E_{B-V} + 3\sigma_{E_{B-V}} < 0$) 11993 из 94670 (12.6%), то есть достаточно много. Основные возможные причины таких отклонений — неверные таблицы «спектральный тип — покраснение», ошибки в спектральной классификации (см. обучающее множество в «Способах получения классов светимости»). Этой проблеме будет посвящена следующая статья.

7 Приложение

7.1 Картирование небесной сферы

В данной работе, все изображения неба являются проекцией Хаммера небесной сферы в галактической системе координат, т.ч. в центре изображения — центр галактики, сверху — северный галактический полюс, снизу — южный галактический полюс, увеличение галактических долгот справа налево.

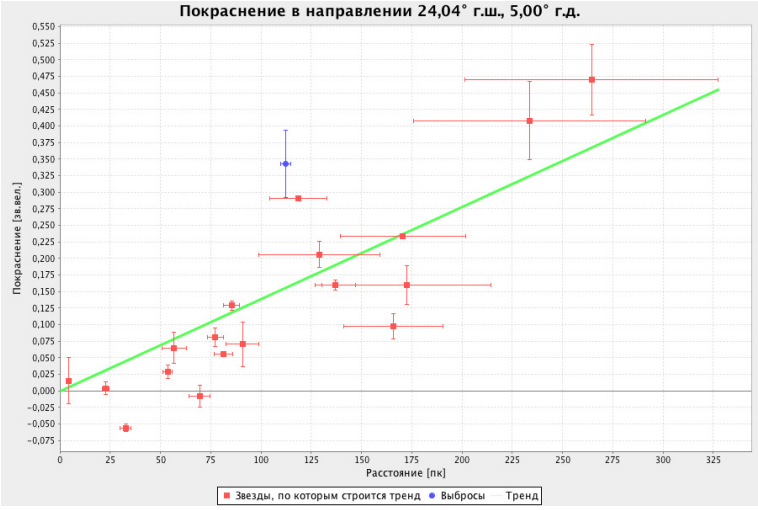

Рис. 8: Тренд покраснения, построенный без учета выбросов

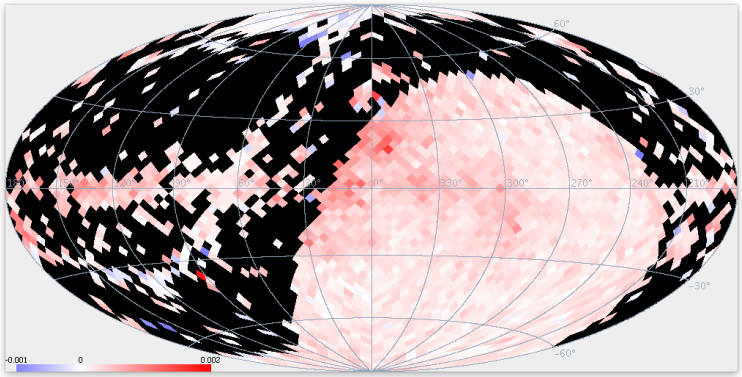

Рис. 9: Распределение градиента покраснения, построенного по звездам, имеющим класс светимости в каталоге Hipparcos и показатель цвета по этому классу светимости в таблице [7] (44658 звезд)

Небесная сфера разбита на «пиксели» стандартным алгоритмом HEALPix [10]. Ключевым параметром (resolution parameter) алгоритма, определяющим разбиение сферы на равные площадки, является число N_{side}. Общее число пикселей $N_{pix} = 12N_{side}^2$. Двумя параллелями со склонениям $\pm\arcsin(2/3)$ вся сфера разбивается на три части — экваториальную и две полярные. В полярных зонах выбирается по $N_{side}-1$ параллелей, в экваториальной зоне число параллелей равно $(2N_{side}+1)$. На каждой параллели экваториальной области находятся центры $4N_{side}$ площадок. Ближайшие к полюсам параллели всегда содержат по четыре площадки, а при движении от полюсов к экватору в полярных зонах число площадок на каждой параллели увеличивается на единицу. Нумерация площадок $i = 0, \ldots, N_{pix} - 1$ идет по параллелям с севера на юг.

7.2 Классификация данных методом опорных векторов

Задача классификации состоит в том, чтобы определить, к какому классу относится данный объект на основе обучающей выборки — других объектов, про которые заранее известно, к каким классам они принадлежат. Каждый объект описывается числовыми атрибутами, поэтому задача классификации

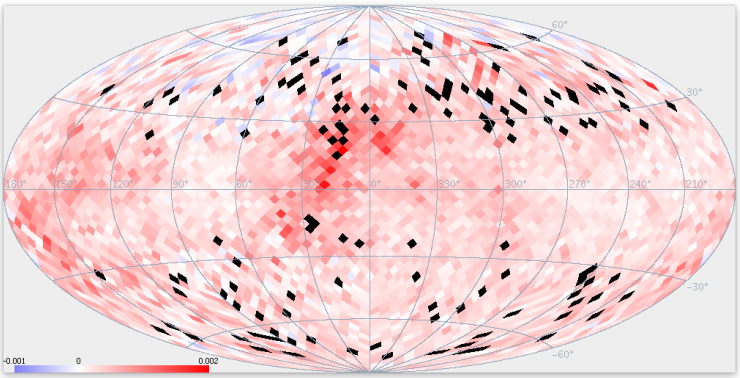

Рис. 10: Распределение градиента покраснения, построенного по звездам с $B - V < 0.6$ (44890 звезд)

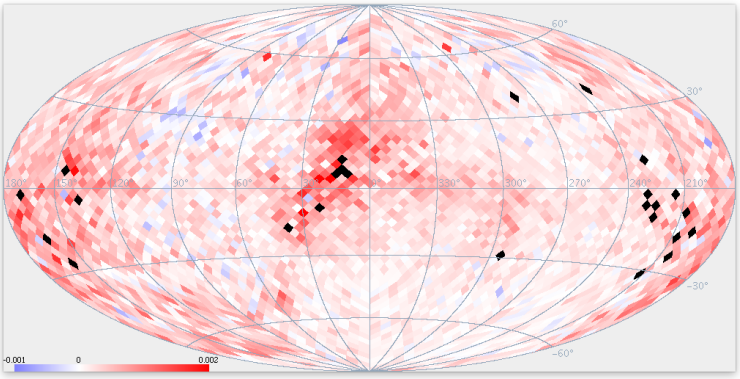

Рис. 11: Распределение градиента покраснения, построенного по звездам с $B - V \geq 0.6$ (49309 звезд)

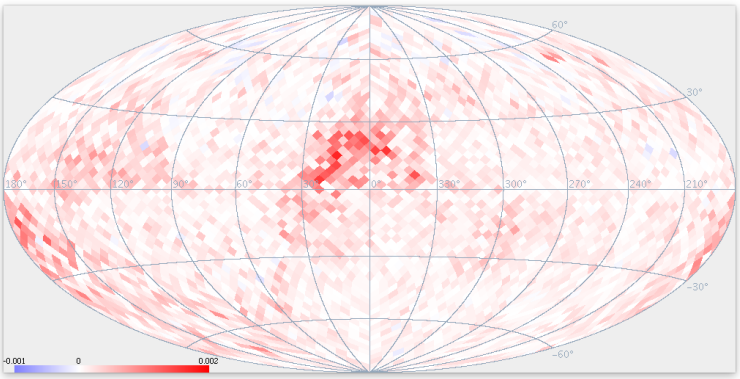

Рис. 12: Распределение градиента покраснения, построенного по всем звездам (94199 звезд)

объектов сводится к задаче классификации точек в $\mathbb{R}^n$. Если классов всего два, то задача называется бинарной классификацией, если несколько — мультиклассификацией.

Задачу классификации на m классов можно сформулировать следующим образом, пусть есть обу-

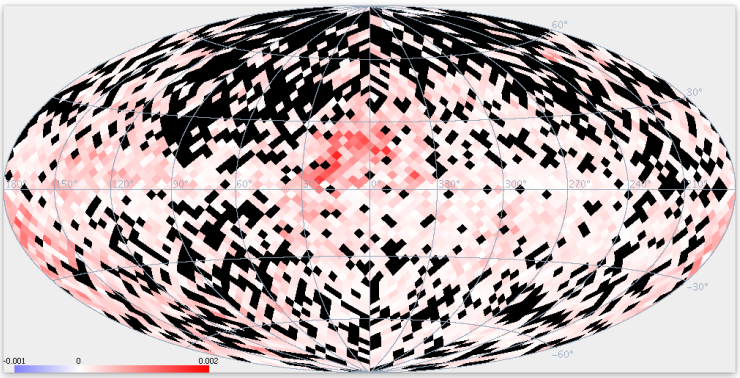

Рис. 13: Карта 12, на которой черным цветом отмечены пиксели, в которых $k < 2\sigma_k$

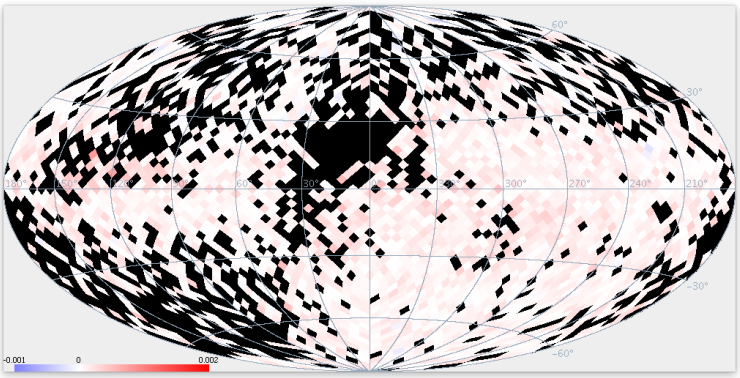

Рис. 14: Карта 12, на которой черным цветом отмечены пиксели, в которых $\sigma_k > 0.05^m/\text{кпк}$

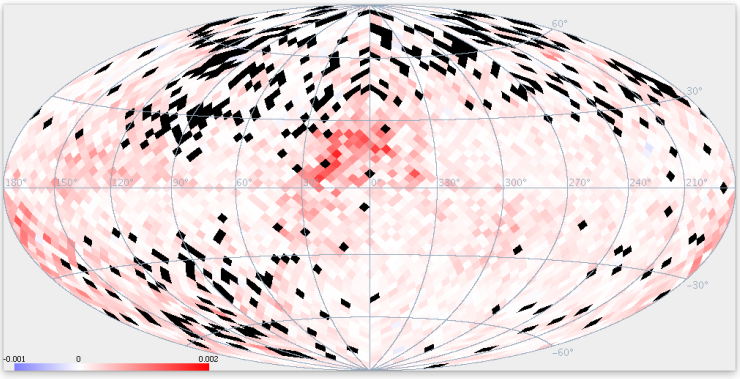

Рис. 15: Карта 12, на которой черным цветом отмечены пиксели, в которых $k < 2\sigma_k$ либо $\sigma_k > 0.05^m/\text{кпк}$

чающая выборка $(x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \{1, \ldots, m\}$. Требуется на основе обучающей выборки построить решающую функцию $F : \mathbb{R}^n \to \{1, \ldots, m\}$, сопоставляющую класс любой точке из $\mathbb{R}^n$. В случае бинар-

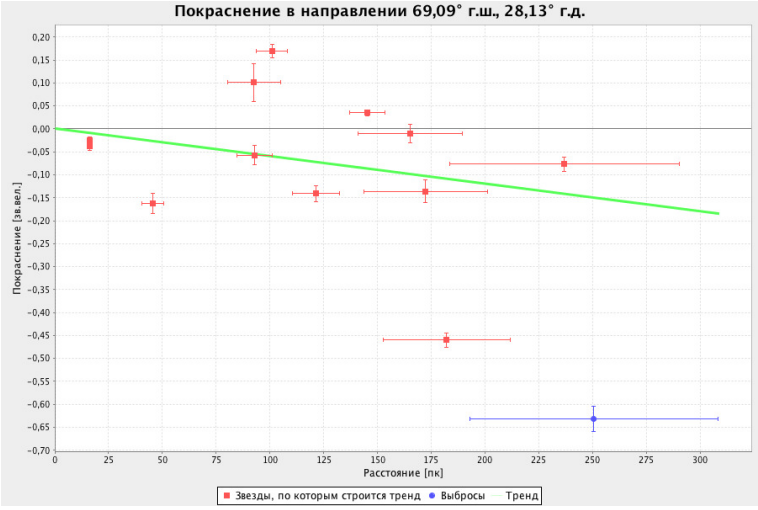

Рис. 16: Отрицательный ход покраснения

ной классификации, классы будет удобно обозначить за $\{-1, 1\}$.

Бинарный классификатор назовем *линейным*, если его решающая функция выглядит следующим образом,

$$F(x) = \text{sign}(w \cdot x + b) = \text{sign}\left(\sum_{i=1}^{n} w_i x_i + b\right),$$

где x_i — компоненты вектора x, b — параметр, w — вектор, соответствующий нормали к *разделяющей гиперплоскости* (ее уравнение $w \cdot x + b = 0$). Классы $\{-1, 1\}$ назовем линейно разделимыми, если существует гиперплоскость, разделяющая точки разных классов по разным полупространствам относительно этой гиперплоскости.

В простейшем случае, метод опорных векторов — это алгоритм обучения линейного бинарного классификатора методом *максимального зазора*. Если классы $\{-1, 1\}$ линейно разделимы, это есть алгоритм нахождения гиперплоскости, которая наилучшим образом разделяет классы. То есть гиперплоскости (w, b), которая находится на максимальном расстоянии до ближайшей точки «-1» класса и ближайшей точки «1» класса. Можно показать, что такая гиперплоскость будет иметь эти расстояния равными $\frac{1}{\|w\|}$ — величине *зазора* (рис 13). Зазор должен быть максимальным, поэтому, для нахождения такой гиперплоскости требуется максимизировать $\frac{1}{\|w\|}$, то есть минимизировать $\|w\|^2$. Тем самым, метод опорных векторов сводится к решению следующей задачи,

$$\begin{cases} \|w\|^2 \to \min \\ \text{sign}(w \cdot x_i + b) = y_i \end{cases}$$

где второе условие соответствует линейной разделимости обучающей выборки. Перепишем,

$$\begin{cases} \|w\|^2 \to \min \\ y_i(w \cdot x_i + b) \geq 1 \end{cases}$$

что есть задача квадратичного программирования, которая решается с помощью множителей Лагранжа.

Для того, чтобы алгоритм мог работать в случае, если классы линейно неразделимы, позволим ему допускать ошибки на обучающей выборке. Введем набор дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибки на точках x_i. Смягчим ограничения в неравенствах и введем в

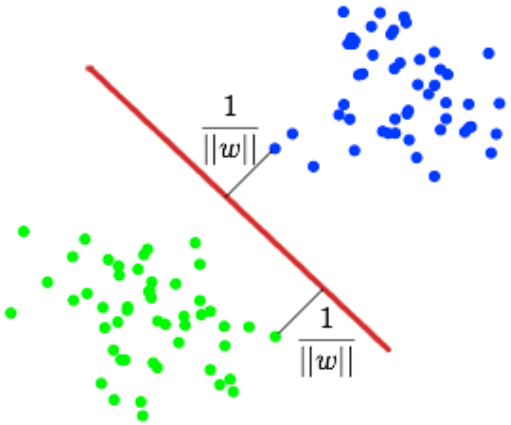

Рис. 17: Решение задачи бинарной классификации в $\mathbb{R}^2$ методом опорных векторов

минимизируемый функционал штраф за суммарную ошибку,

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \to \min \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Где C — параметр настройки метода. Опять же, получилась задача квадратичного программирования. Эффективно такая задача может решаться такая с помощью метода «Sequential Minimal Optimization» (SMO) [9].

7.3 Кросс-валидация, оценка качества работы классификатора

Для получения несмещенной оценки качества работы аналитической модели проводится *кросс-валидация*. Она заключается в том, что обучающая выборка дизъюнктно делится на две части — тренировочное множество и тестовое множество, затем модель обучается на тренировочном множестве, а оценка качества работы получается на тестовом. Одним из способов проведения кросс-валидации является k-fold кросс-валидация. В этом случае обучающая выборка разбиваются на k частей, затем на $k-1$ части выборки производится обучение модели, а оставшаяся часть используется для тестирования. Процедура повторяется k раз, в итоге каждая из k частей используется в качестве тестовой. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Для задачи классификации на m классов, результатом кросс-валидации является таблица $m \times m$, в ячейке i, j которой записано число объектов класса i из тестового множества, про которые классификатор считает, что они принадлежат классу j. В случае k-fold кросс-валидации итоговая таблица $m \times m$ есть сумма всех k таблиц, полученных из обучения + тестирования классификатора на каждом из k разбиений на тренировочное и тестовое множество.

По полученной таблице $C[m, m]$ можно считать различные метрики — числа, характеризующие качество классификации,

- *Точность* определения класса i — $\frac{C_{ii}}{\sum\limits_{j=1}^{m} C_{ji}}$

- *Полнота* определения класса i — $\frac{C_{ii}}{\sum_{j=1}^{m} C_{ij}}$
- F_1-*мера* определения класса i - среднее гармоническое точности и полноты определения класса i

8 Заключение

Этот раздел еще не создан

Список литературы

[1] Perryman M.A.C., Lindegren L., Kovalevsky J., Hog E., Bastian U., Bernacca P.L., Creze M., Donati F., Grenon M., Grewing M., van Leeuwen F., van der Marel H., Mignard F., Murray C.A., Le Poole R.S., Schrijver H., Turon C., Arenou F., Froeschle M., Petersen C.S., "The Hipparcos Catalogue"(1997A&A...323L..49P)

[2] Hog E., Baessgen G., Bastian U., Egret D., Fabricius C., Grossmann V., Halbwachs J.L., Makarov V.V., Perryman M.A.C., Schwekendiek P., Wagner K., Wicenec A., "The Tycho Catalogue"(1997A&A...323L..57H)

[3] van Leeuwen F., Evans D.W., Grenon M., Grossmann V., Mignard F., Perryman M.A.C., "The Hipparcos mission: photometric data."(1997A&A...323L..61V)

[4] Hipparcos, the New Reduction of the Raw Data van Leeuwen F.,Astron. Astrophys. 474, 653 (2007),

[5] Wright et al. *Tycho-2 Spectral Type Catalog* 2003: The Astronomical Journal

[6] Страйджест – книжка

[7] А.А.Сминов, А.С.Цветков, А.В.Попов *Неточности в спектальной классификации звезд каталога Tycho-2 Spectral Type* 2006.

[8] V.N.Vapnik *The Nature of Statistical Learning Theory* 1995.

[9] Hastie, T.; Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning* , Springer New York Inc. , New York, NY, USA .

[10] Gorski et al. *HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere* 2005.